

Constructing an Adaptive Learning Framework Integrating STEAM–Gamification for Numeracy and Computational Thinking in Remote Indonesian Elementary Schools

Masnur*, Nadar, & Saleha

Department of Primary Education, Universitas Muhammadiyah Enrekang, Indonesia

*Corresponding email: masnur@unimen.ac.id

Received: 03 May 2026

Accepted: 17 September 2026

Published: 10 October 2026

Abstract: This study aimed to construct and validate the design of an adaptive learning framework that integrates STEAM–gamification principles with Zone of Proximal Development (ZPD)–Item Response Theory (IRT) to simultaneously develop numeracy literacy and computational thinking among elementary school students in Indonesia’s 3T (remote, frontier, underdeveloped) regions. A mixed-methods sequential exploratory design integrated with Design-Based Research (DBR) was applied over eight months in Pinrang Regency, South Sulawesi, across 10 schools in the qualitative phase and 20 schools in the quantitative phase. Qualitative data were drawn from interviews and focus group discussions with 30 teachers, 10 principals, 30 students, and 20 stakeholders, alongside 30 classroom observations; quantitative data were collected from 300 Grade IV–V students and 60 teachers using structured surveys, IRT-calibrated diagnostic tests (3PL; $\alpha = .83$), and an ICT infrastructure audit. The framework synthesized qualitative and quantitative findings into four integrated layers: Pedagogical (STEAM), Motivational (gamification), Adaptive (ZPD–IRT), and Assessment (computational thinking + numeracy). Expert design validation employed LORI, Aiken’s V, and Fleiss’ Kappa with eight domain specialists. The needs analysis showed that 85% of teachers relied exclusively on conventional instruction, 78% identified a strong need for interactive learning media, and only 30% of schools had adequate ICT infrastructure. The composite numeracy mean was 50.1/100 and the computational thinking mean was 48.3/100, with geometry the weakest numeracy domain (34.7). Economic feasibility yielded a benefit-cost ratio of 1.42, a positive NPV, and an internal rate of return of 18.3%. Expert design validation yielded a LORI mean of 4.38/5, a Fleiss’ Kappa of 0.74, and an Aiken’s V of 0.79, all exceeding acceptability thresholds. The four-layer adaptive framework, among the first to integrate STEAM, gamification, ZPD–IRT adaptive learning, and computational thinking as a coherent system for Indonesia’s 3T elementary schools, provides a replicable theoretical model and an expert-validated design basis for subsequent product development and efficacy trials

Keywords: adaptive learning framework, STEAM–gamification, numeracy literacy, computational thinking, 3T elementary schools.

Article’s DOI: <https://doi.org/10.23960/jpp.v16i4.pp1993-2008>

■ INTRODUCTION

Numeracy literacy refers to the ability to understand and apply mathematical concepts to solve real-world problems across diverse domains (Heryani et al., 2023, p. 2; Susandi et al., 2025). Its significance for human capital development in the twenty-first century has prompted increased research and policy attention

to the persistent gap between educational objectives and actual student performance, particularly in countries with unequal educational resources. Indonesia’s performance on international assessments underscores this disparity: in the 2022 PISA cycle, Indonesia ranked 75th out of 80 countries, with an average mathematics score of 379, nearly 100 points

below the OECD average of 472 (Jusniani et al., 2023). Nationally, only 40.63% of junior secondary students met the minimum numeracy standard in the 2023 National Assessment (Silvia & Asdarina, 2024), and about 70% of elementary students showed significant weaknesses in geometry (Nurhayati et al., 2022). The 2025 Academic Competency Test (TKA) reported a national mathematics average of 36.1, with South Sulawesi Province at 34.67, further illustrating the issue's severity and pervasiveness.

Persistent numeracy challenges stem not only from resource limitations but also from fragmented learning frameworks. Approaches such as STEAM (Science, Technology, Engineering, Arts, Mathematics), gamification, adaptive learning, and computational thinking have each demonstrated effectiveness in enhancing student engagement and achievement in mathematics (Budiyono et al., 2023; Kabilan et al., 2023; Zhang & Huang, 2024). However, these strategies are frequently implemented in isolation rather than as components of an integrated, cohesive system. This fragmentation is also evident in the Indonesian elementary context specifically: STEAM-based media for young learners have been developed without adaptive or motivational components (Puspitasari et al., 2022; Maulidiansy et al., 2023), interactive numeracy multimedia has been produced without gamified or adaptive features (Salsabila et al., 2022), and gamification in STEAM has so far been examined mainly at the level of student perception rather than as part of an integrated learning system (Budiyono et al., 2023). Consequently, some interventions prioritize enjoyment at the expense of rigor, while others offer personalization without adequately addressing students' cultural or motivational contexts. Wing (2006) emphasized that computational thinking, encompassing skills such as problem decomposition, pattern recognition, abstraction, and algorithmic reasoning, is essential

for the twenty-first century. Nevertheless, Indonesian elementary education has not systematically incorporated computational thinking into numeracy assessment.

These challenges are exacerbated in Indonesia's 3T regions (Terdepan, Terluar, Tertinggal, referring to frontier, outermost, and underdeveloped areas), where inadequate ICT infrastructure, limited teacher digital competencies, and restricted access to contemporary teaching resources intensify educational disparities. Kabupaten Pinrang in South Sulawesi exemplifies these issues; in its 3T sub-districts, such as Kecamatan Patampanua, both internet and computer access remain scarce. Preliminary observations indicate that 85% of elementary teachers still rely primarily on textbooks, and only 30% of schools have functional computer laboratories with reliable internet connectivity. These circumstances underscore both the urgent need for adaptive learning solutions and the significant obstacles to their implementation.

A review of existing literature identifies four principal gaps that this study seeks to address. First, interactive multimedia for numeracy (Salsabila et al., 2022) and STEAM-based instructional materials for young learners (Puspitasari et al., 2022; Maulidiansy et al., 2023) have been developed without incorporating adaptive features or conducting thorough needs assessments. Second, many internationally designed adaptive learning tools presuppose infrastructural conditions that are not present in 3T regions (Kalogiannakis et al., 2021). Third, gamification and adaptive learning are frequently examined independently rather than as an integrated approach (Chen et al., 2023; David & Weinstein, 2023). Fourth, no prior research in Indonesia has addressed numeracy literacy and computational thinking as concurrent objectives within a unified adaptive framework. These gaps highlight the need for research that constructs a

practical, replicable framework, moving beyond mere problem identification.

Building on these gaps, this study constructs and validates a four-layer adaptive learning framework, covering pedagogy (STEAM), motivation (gamification), adaptation (ZPD–IRT), and assessment (computational thinking and numeracy). To make the research problem explicit, the study is guided by the following research questions (RQs):

- RQ1. What are the numeracy and computational thinking learning needs of students and teachers in Pinrang's 3T elementary schools, and what field conditions constrain them?
- RQ2. To what extent is an adaptive STEAM–gamification framework technically, pedagogically, and economically feasible in this low-resource setting?
- RQ3. How can mixed-methods needs evidence be synthesized into a coherent four-layer adaptive learning framework (pedagogical, motivational, adaptive, and assessment) suited to 3T constraints?
- RQ4. To what degree do domain experts judge the resulting framework design to be valid, in terms of content validity (Aiken's V, CVR), quality (LORI), and inter-rater agreement (Fleiss' Kappa)?

By developing and validating this framework design in a 3T region, the study offers both a theoretical model for adaptive learning in resource-limited settings and a solid, expert-validated foundation for the product-development and efficacy-testing phases that follow.

■ METHOD

Research Design

This study used a mixed-methods, sequential, exploratory design (Creswell & Plano Clark, 2018) combined with Design-Based Research (DBR) to construct and validate an adaptive learning framework. DBR was selected because it builds and refines models from

authentic educational practice through iterative cycles (Anderson & Shattuck, 2012). The research ran over eight months and was organized into four sequential DBR phases, each informing the next: (1) needs analysis and theoretical construct mapping (qualitative); (2) feasibility study and baseline measurement (quantitative); (3) framework construction (integration and synthesis); and (4) expert validation and iterative revision. The sub-sections below report participants, instruments, ethical considerations, and analysis across these phases.

Participants and Sampling

This study took place in Kabupaten Pinrang, South Sulawesi Province, which has 287 public elementary schools, approximately 18,734 students, and 1,245 teachers; several sub-districts are classified as 3T (frontier, outermost, and disadvantaged) areas. In the qualitative phase, purposive sampling was used to select 10 schools. To make the sampling replicable, schools were selected against explicit inclusion criteria: (a) geographic stratum (e" 3 urban/peri-urban and e" 4 3T/remote schools, classified using the village Developing/Underdeveloped status in the official Indeks Desa Membangun); (b) accreditation balance across grades A, B, and C; and (c) ICT-infrastructure tier defined a priori as High (functional computer lab + stable grid electricity + fixed/broadband internet), Medium (shared/limited devices + intermittent connectivity via mobile data), or Low (no lab, reliance on teacher-owned smartphones, frequent power interruptions). At least two schools were drawn from each ICT tier so that the full infrastructure spectrum was represented. For the quantitative phase, 20 schools participated, involving 300 Grade IV–V students and 60 teachers.

Ethical Considerations

Because the study engaged children and teachers as human participants, ethical safeguards were applied throughout. Ethical clearance was

obtained from the institutional research ethics committee of Universitas Muhammadiyah Enrekang, and research permits were issued by the Pinrang District Education Office and each participating school. Written informed consent was obtained from teachers, principals, and stakeholders; for students, written parental/guardian consent together with child assent was secured prior to participation. Participation was voluntary, data were anonymized and stored securely, and participants could withdraw at any time without consequence.

Data Collection Instruments

Phase 1 (qualitative, months 1–2) used four techniques: in-depth semi-structured interviews with 30 teachers, 10 principals, and 30 students; focus group discussions (FGDs) with 20 education stakeholders (district supervisors, curriculum specialists, school-committee representatives); 30 structured classroom observations documenting lesson structure, student behavior, media use, and facility constraints; and document analysis of curriculum plans, lesson plans, and assessment records.

In parallel, a systematic literature review was conducted following the PRISMA 2020 reporting guidelines to map research on STEAM, gamification, differentiated/adaptive learning, and computational thinking. The full search strategy (databases: Scopus, ERIC, and Garuda; Boolean strings; 2015–2025 window; English/Indonesian), the inclusion/exclusion criteria, and the PRISMA flow diagram (records identified, screened, excluded, and included) are provided in Appendix A to substantiate the PRISMA claim. To avoid methodological confusion, the review is positioned as a supporting construct-mapping activity that informed framework design, not as the study's primary empirical method.

Phase 2 (quantitative, months 3–5) used four instruments. The Student Numeracy and Computational Thinking Diagnostic Test

(SNCTDT) contained 40 numeracy and 25 computational thinking items adapted from PISA 2022, TIMSS 2023, and the Bebras Challenge, covering number, algebra, geometry, measurement, data, decomposition, pattern recognition, abstraction, and algorithmic thinking. Item fit ($RMSEA < .06$) and reliability ($\alpha = .83$ for numeracy; $\alpha = .80$ for computational thinking) were confirmed using IRT 3PL calibration in BILOG-MG; marginal (IRT-based) reliability and the test information function were additionally examined. Given the sample size ($n = 300$), interpret 3PL estimates, particularly the guessing parameter, cautiously and re-estimate them with a larger sample in subsequent phases.

To prevent cognitive fatigue from administering 65 items to Grade IV–V children, the SNCTDT was delivered in two proctored 35-minute sessions on separate days (Session 1: 40 numeracy items; Session 2: 25 computational thinking items), each preceded by a 5-minute worked example and a short break, with trained field facilitators reading instructions aloud to control for reading-fluency effects.

The Teacher Readiness Survey (TRS) was a 45-item Likert instrument measuring TPACK competency and digital literacy (reliability $\alpha = .87$). Construct validity of the TRS was established before use: content validity via expert review (item-level Aiken's $V = .78-.92$), and construct adequacy via exploratory factor analysis ($KMO = .87$; Bartlett's test $p < .001$; five factors with eigenvalues > 1 explaining 63.4% of variance; item factor loadings $r = .42-.79$, all above the .40 retention threshold).

The ICT Infrastructure Audit Checklist recorded device availability, device types, connectivity stability, and electrical supply across 20 schools. Economic feasibility was assessed through a cost-benefit analysis, net present value (NPV), benefit-cost ratio (BCR), internal rate of return (IRR), payback period, and a sustainability

index, computed over a five-year horizon at a 12% discount rate. The monetization basis is reported transparently in the Data Analysis subsection and in Appendix B (cost components, benefit-valuation logic, and a one-way sensitivity analysis).

Data Analysis

Qualitative data were analyzed thematically in NVivo 14 using Braun and Clarke's (2006) six-phase approach, triangulated with Miles and Huberman's (1994) interactive analysis model. Concretely, transcripts were processed through data condensation (open coding of meaning units), followed by axial coding to cluster codes into categories, then data display in thematic matrices and a thematic map, and finally conclusion drawing and verification through member checking and inter-coder agreement (two coders; Cohen's $\kappa = .81$).

Quantitative analyses comprised IRT 3PL calibration of the diagnostic items, descriptive statistics for domain/sub-construct means, and the economic indicators above. For the economic model, benefits were monetized from three transparent sources: (a) avoided remediation cost (number of at-risk students $\times$ unit cost of district remedial programs), (b) teacher time-efficiency gains (hours saved $\times$ teacher hourly wage proxy), and (c) projected reduction in grade-repetition cost (repetition-rate change $\times$ per-student annual cost); costs included development, devices, training, content, and maintenance. A one-way sensitivity analysis varied the discount rate (8–16%) and the benefit-realization rate ($\pm 25\%$) to test robustness. Appendix B itemizes all monetary parameters and their sources.

For expert validation, the framework and blueprint were rated by eight specialists on the Learning Object Review Instrument (LORI; Nesbit et al., 2009) across nine quality dimensions (five-point scale). Content validity was quantified with Aiken's V (threshold $e^* .75$) and the Content Validity Ratio (CVR). Because LORI ratings are ordinal, inter-rater reliability was summarized with Fleiss' Kappa (threshold $e^* .60$) and complemented by the intraclass correlation coefficient (ICC) and Kendall's W . Internal consistency of the diagnostic instrument was verified via Cronbach's α (threshold $e^* .80$). A minimum of two iterative revision cycles were completed before the framework design was declared validated.

RESULT AND DISCUSSION

Consistent with the four DBR phases, the findings are organized into four thematic subsections: (a) the numeracy and computational thinking needs profile and field conditions; (b) baseline competency and multidimensional feasibility; (c) framework characteristics and expert validation; and (d) practicality indicators, effectiveness scope, and the roadmap to efficacy testing. Because this paper reports the design-and-validation stage of an R&D program, the discussion of "effectiveness" is deliberately framed as planned efficacy work rather than as realized learning-gain claims (see Limitations).

Needs Profile and Field Conditions (RQ1)

To visualize how these field conditions interact rather than operate in isolation, Figure 1 presents a thematic map synthesizing the Phase 1 qualitative findings.

Table 1. Needs profile of numeracy learning in 3T elementary schools, pinrang regency

Dimension	Key Finding	Evidence Source	Implication for Media Design
Pedagogical Practice	85% of teachers rely on textbook-centered instruction; activity-based approaches in <15% of observed lessons	Classroom observation (n=30); teacher interviews	Embed active, problem-based numeracy tasks alongside traditional methods

Student Motivation	72% of students reported low intrinsic motivation; 81% preferred game-like activities	Student interviews; FGD; off-task observation	Use theory-grounded gamification (points, badges, narrative)
ICT Infrastructure	Only 30% have functional labs; 45% have erratic electricity; 62% rely on mobile data.	ICT Audit Checklist (n=20 schools)	Offline-first, optimized for low-end Android (32-bit, ≤2GB RAM)
Teacher Capacity	78% need interactive numeracy media; mean digital literacy = 2.61/5 (below 3.5)	TRS survey; principal interviews	Teacher-accessible UI; minimal training; built-in pedagogical guides
Community Aspiration	Principals/supervisors supported STEAM-gamification; parents accepted supervised use.	FGD (20 stakeholders); principal interviews	Time-bounded sessions; parent-guidance modules

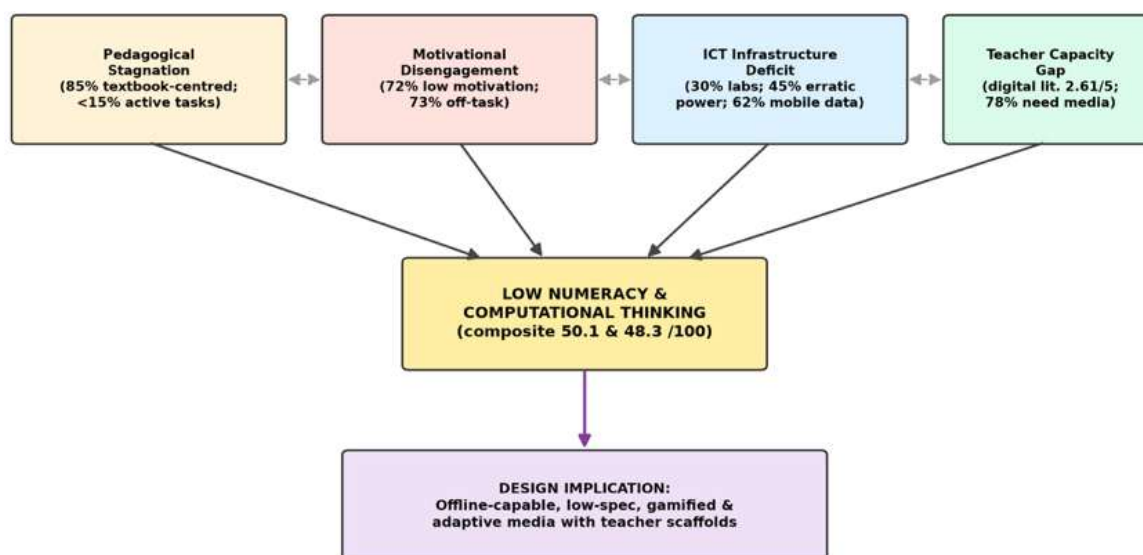


Figure 1. Thematic map of field conditions constraining numeracy literacy in 3T schools

Pedagogical stagnation was evident across all sites: lessons followed a uniform pattern of textbook worked-examples, copying procedures, and routine practice, with little discovery-based, hands-on, or technology-supported activity. These observations echo Putrawangsa and Hasanah (2022), who linked Indonesian students' low PISA scores to procedural memorization rather than conceptual understanding, and Salsabila et al. (2022), who showed that interactive multimedia can broaden available teaching methods.

Motivational disengagement was visible in 73% of observed sessions (off-task, passive, low

participation); students described mathematics as “dull” and “hard to understand without real-life examples.” This supports Self-Determination Theory (Ryan & Deci, 2000), and well-designed gamification can satisfy the underlying needs for autonomy, competence, and relatedness (David & Weinstein, 2023; Zhang & Huang, 2024); Budiyo et al. (2023) similarly reported higher motivation in STEAM-gamification settings. Although 68% of students owned smartphones at home, schools had minimal computing resources, unreliable electricity, and frequently no internet, conditions that, as Kalogiannakis et al. (2021) note, exceed what many adaptive tools

assume. Offline, low-bandwidth media are therefore a necessity, not a convenience, for equitable access.

Baseline Competency and Multidimensional Feasibility (RQ2)

Phase 2 data calibrated the adaptive layer. The SNCTDT yielded a mean numeracy score

of 50.1/100 ($SD = 12.6$) and a mean computational thinking score of 48.3/100 ($SD = 13.1$), both below competency thresholds. As shown in Table 2, geometry was the weakest numeracy domain, and as shown in Table 4 and Figure 2, abstraction and algorithmic thinking were the weakest computational thinking sub-constructs.

Table 2. Numeracy diagnostic test results by domain (n = 300 students)

Numeracy Domain	Grade IV Mean	Grade V Mean	Overall Mean	SD	Category
Number & Operations	52.4	57.1	54.8	11.2	Low
Algebra	48.6	53.9	51.3	12.8	Low
Geometry	31.2	38.2	34.7	14.3	Very Low
Measurement	49.8	54.6	52.2	11.9	Low
Data & Statistics	44.1	49.3	46.7	13.1	Low
Composite Score	47.3	52.8	50.1	12.6	Low

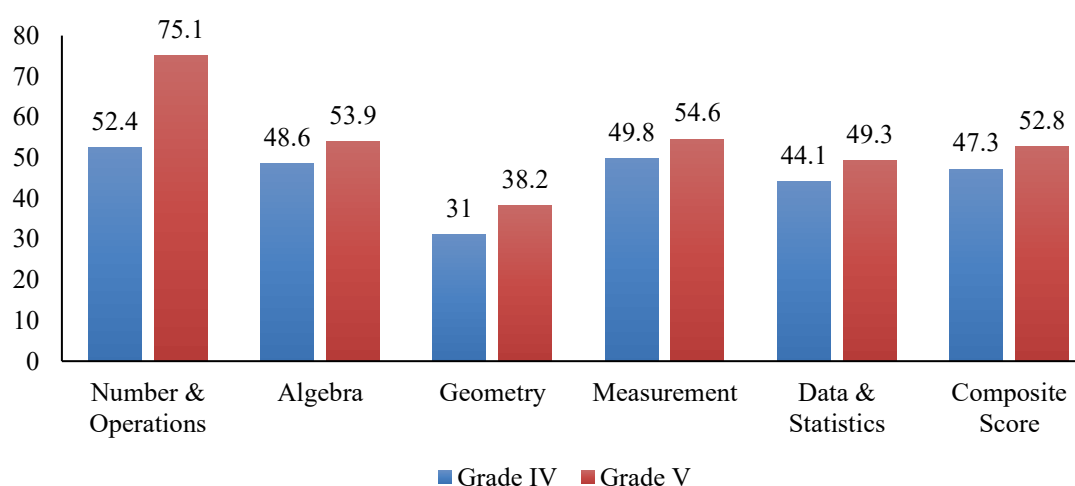


Figure 2. Numeracy diagnostic results by domain and grade

Table 2 presents the numeracy diagnostic results for Grade IV and Grade V students across five domains. Ranked by overall mean, students performed best in Number and Operations ($M = 54.8$, $SD = 11.2$), followed by Measurement ($M = 52.2$, $SD = 11.9$), Algebra ($M = 51.3$, $SD = 12.8$), and Data and Statistics ($M = 46.7$, $SD = 13.1$). These four domains were classified as Low. Geometry had the lowest mean score ($M = 34.7$, $SD = 14.3$) and was the only

domain classified as Very Low. This finding is consistent with the national trend reported by Nurhayati et al. (2022). Students may find Geometry challenging because it involves using area and perimeter formulas, reasoning about space, and applying concepts to real-world situations. The results may also reflect limited opportunities for visual and spatial activities in typical classroom instruction. Scores in Geometry were the most variable of all domains,

although *SD* values differed only modestly across domains (11.2 to 14.3).

Grade V students had higher descriptive scores than Grade IV students in every domain. The smallest difference was in Number and Operations (4.7 points), while the largest was in Geometry (7.0 points). Because the data came from two different grade-level groups at one point in time (a cross-sectional design), these differences should not be interpreted as individual student growth. Further statistical testing is needed to determine whether these differences are significant. Overall, the findings highlight the need

to strengthen numeracy instruction, especially in Geometry. Activities involving visualization, hands-on work with shapes, and problem-solving in real-world contexts may help address this need.

Table 3 summarizes the computational thinking diagnostic results for Grade IV and Grade V students across four sub-constructs. Overall, students were in the Low achievement category, with a composite score of 48.3 (*SD* = 13.1). Grade IV students scored 46.4, compared with 49.8 for Grade V students, a difference of 3.4 points.

Table 3. Computational thinking diagnostic results by sub-construct

CT Sub-Construct	Grade IV Mean	Grade V Mean	Overall Mean	SD	Category
Decomposition	52.1	55.8	54.2	12.4	Low
Pattern Recognition	49.7	53.2	51.6	12.9	Low
Abstraction	42.0	45.3	43.8	13.5	Very Low
Algorithmic Thinking	41.9	45.0	43.6	13.6	Very Low
Composite Score	46.4	49.8	48.3	13.1	Low

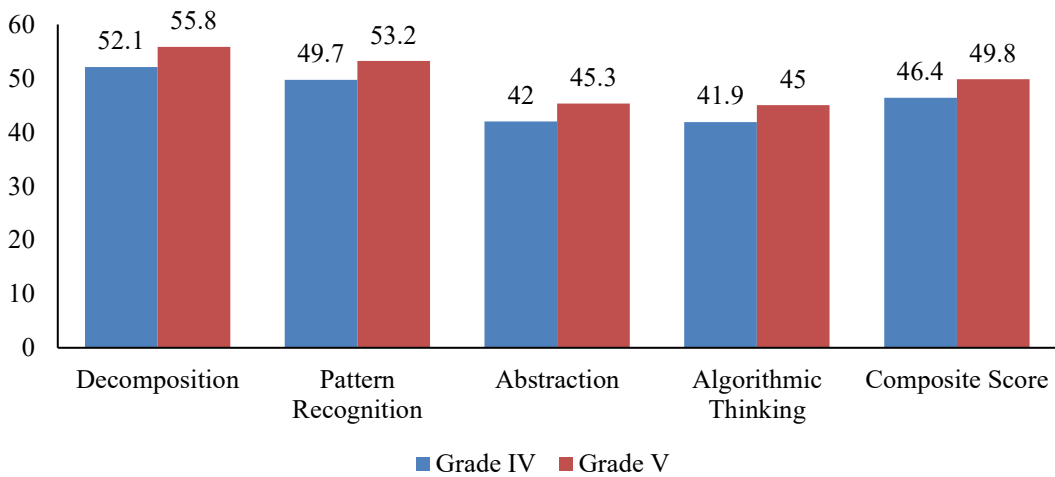


Figure 3. Computational thinking results by sub-construct and grade

Overall, students scored highest in Decomposition (*M* = 54.2, *SD* = 12.4), followed by Pattern Recognition (*M* = 51.6, *SD* = 12.9). Both sub-constructs were in the Low category. Scores were lower in Abstraction (*M* = 43.8, *SD* = 13.5) and Algorithmic Thinking (*M* = 43.6,

SD = 13.6), which were both in the Very Low category. The two scores were very close, differing by just 0.2 points. Both were also below Pattern Recognition: Abstraction by 7.8 points and Algorithmic Thinking by 8.0 points. This pattern suggests that students found it easier to

break problems into parts and recognize patterns than to identify key information and organize steps to solve a problem. The standard deviations ranged from 12.4 to 13.6, indicating similar variation in scores across the four sub-constructs.

Grade 5 students scored higher than Grade 4 students in all four sub-constructs, with differences ranging from 3.1 to 3.7 points. However, the students were in different grade-level groups, and the data were collected at a single point in time. These differences therefore do not show individual student growth. Further statistical testing is needed to determine whether they are significant. Overall, the results suggest that computational thinking instruction could be strengthened, especially in Abstraction and Algorithmic Thinking.

The teacher survey showed mean digital literacy of 2.61/5 and mean TPACK of 2.74/5 (both below 3.5); only 22% had used digital media in numeracy lessons, yet 78% expressed a need for interactive numeracy media, a gap addressable through structured training. Feasibility was confirmed across three dimensions. Pedagogically, test and qualitative data converged on a significant, addressable learning gap, and the expert panel’s first review found alignment with constructivist principles and Universal Design for Learning (UDL). Technically, 68% of students had Android-capable smartphones, and the district office endorsed a two-year phased infrastructure plan. As presented in Table 4, the economic analysis was favorable.

Table 4. Economic feasibility analysis of the adaptive STEAM-Gamification design

Parameter	Value	Interpretation	Verdict
Net Present Value (NPV)	IDR 842,600,000	Positive	Feasible
Benefit-Cost Ratio (BCR)	1.42	Benefits exceed costs by 42%	Feasible
Internal Rate of Return (IRR)	18.3%	Above discount rate (12%)	Feasible
Payback Period	3.2 years	Within project horizon	Feasible
Sustainability Index	0.74 (scale 0–1)	High long-term viability	Feasible

A BCR of 1.42 and a positive NPV indicate that projected social and educational benefits exceed development and implementation costs. To make these figures transparent (rather than appearing ad hoc), the NPV of IDR 842,600,000 aggregates three monetized benefit streams avoided remediation cost, teacher time-efficiency gains, and reduced grade-repetition cost net of development, device, training, content, and maintenance costs, discounted at 12% over five years; a one-way sensitivity analysis (discount rate 8–16%; benefit realization ±25%) is reported in Appendix B and keeps the BCR above 1.0 across all tested scenarios. This aligns with evidence on the economic rationale for educational technology in developing contexts (Angrist et al., 2023, p. 70), especially with scalable low-cost platforms.

Framework Characteristics and Expert Validation (RQ3–RQ4)

Integrating Phase 1 and Phase 2 findings produced the four-layer Adaptive Learning Framework (ALF-STEAM-G). Figure 1 presents the overall architecture and inter-layer relationships, and Figure 6 traces how individual learner data flow through the four layers (entry → .STEAM scenario → gamified motivation → ZPD–IRT routing → dual assessment → 0 feedback loop), responding to the request for a system-level flow representation. Table 6 then summarizes each layer’s theoretical basis, components, empirical input, and 3T adaptation.

The most distinctive theoretical contribution lies in Layer 3. Unlike server-dependent Computerized Adaptive Testing (Boussakuk et

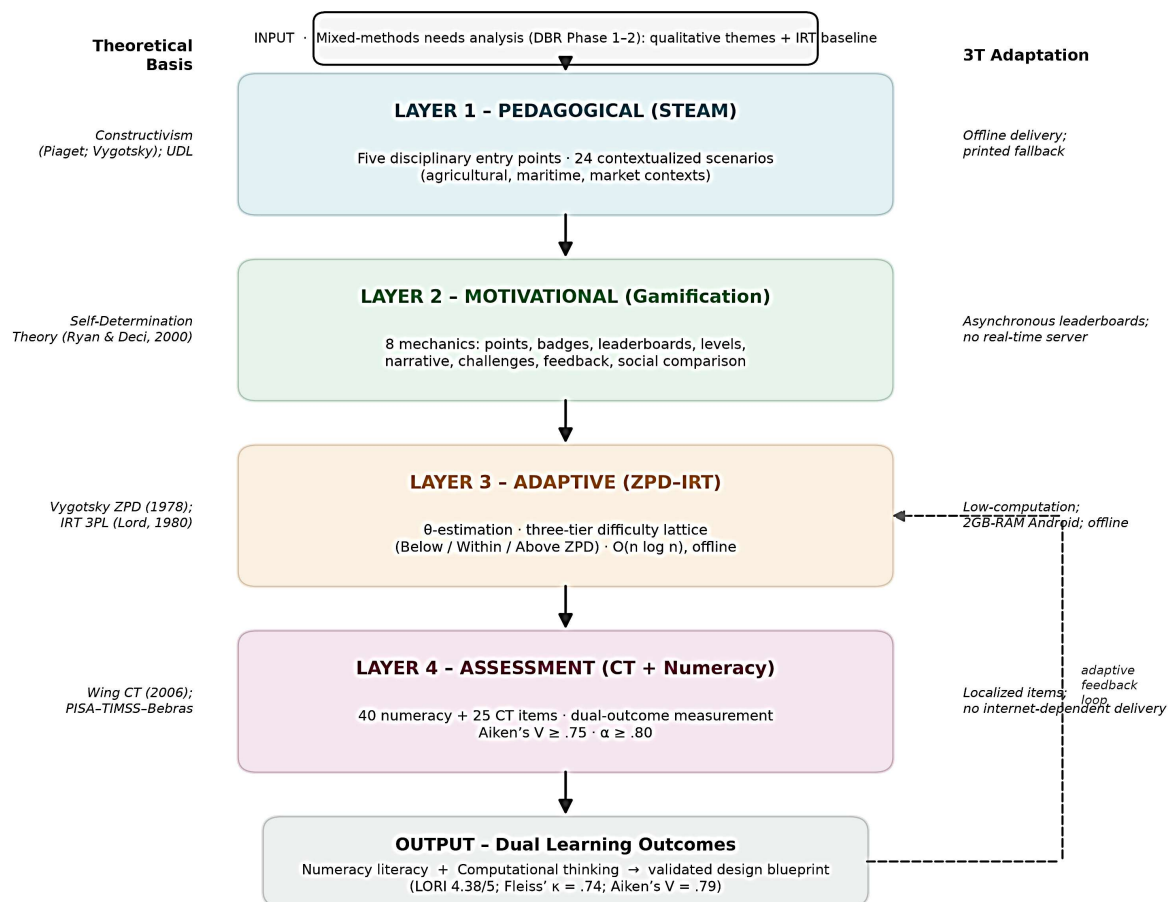


Figure 4. Learner data-flow diagram across the four integrated layers

Table 5. Architecture of the four-layer adaptive learning framework (ALF-STEAM-G)

Layer	Theoretical Basis	Key Components	Empirical Input	3T Adaptation
1. Pedagogical (STEAM)	Constructivism (Piaget & Vygotsky); UDL	Five disciplinary entry points and twenty-four contextualized scenarios derived from local agricultural, maritime, and market environments.	15 thematic needs; lesson observation data	Offline scenario delivery; printed fallback materials
2. Motivational (Gamification)	Self-Determination Theory (Ryan & Deci, 2000)	8 mechanics: points, badges, leaderboards, levels, narrative, challenges, feedback, social comparison	72% low intrinsic motivation; student preference data (FGD)	Asynchronous leaderboards; no real-time server required
3. Adaptive (ZPD-IRT)	Vygotsky ZPD (1978); IRT 3PL (Lord, 1980)	Theta (θ) Estimation, Three-Tier Difficulty Lattice (Below, Within, Above ZPD), and $O(n \log n)$ Algorithm	Baseline θ distribution; IRT item bank calibrated from diagnostic test	Low-computation design; 2GB RAM Android-compatible; full offline capability
4. Assessment (CT + Numerasi)	Wing CT Framework (2006); PISA-TIMSS-Bebras	40 numeracy + 25 CT items; dual outcome measurement; Aiken's $V \geq 0.75$; $\alpha \geq 0.80$	SNCTDT calibration data; expert content review	Contextualized items using local problem scenarios; no internet-dependent delivery

al., 2021, p. 115), the ZPD–IRT algorithm uses an efficient è-estimation procedure over a locally stored, pre-assembled item lattice, eliminating real-time server access. Algorithmically, the design targets $O(n \log n)$ complexity and is intended to run on Android devices with as little as 2GB RAM; however, because no full prototype field deployment has yet been benchmarked, this remains a design specification rather than a verified runtime result. A controlled device-performance test (load time, memory footprint, battery use across low- and mid-tier devices) is planned for the product-development phase, and the corresponding performance table will report empirical values. Zhang and Huang (2024) similarly showed adaptive gamified assessment can yield gains without high-specification infrastructure; the present design

extends this toward a fully offline-capable model.

The explicit integration of computational thinking as a co-equal outcome in Layer 4 is a second innovation. To our knowledge, no prior published study in the Indonesian elementary context has operationalized both numeracy and computational thinking as dual outcomes within one adaptive framework. Wing (2006) frames computational thinking as a universal competency that intersects mathematical reasoning, and the Bebras Challenge (Dagiene & Futschek, 2008) translates these constructs into age-appropriate tasks. Localization via Pinrang scenarios (rice-field area, fish-catch data, market pricing) supports content validity and cultural relevance, addressing the contextual-relevance need highlighted by Chen et al. (2023).

Table 6. Expert validation results for ALF-STEAM-G framework using LORI instrument (n = 8 experts)

LORI Quality Dimension	Mean Score (/5)	SD	CVR	Category
1. Learning Goal Alignment	4.50	0.38	0.87	Excellent
2. Feedback and Adaptation	4.38	0.52	0.82	Very Good
3. Motivation	4.63	0.35	0.91	Excellent
4. Presentation Design	4.25	0.46	0.78	Very Good
5. Interaction Usability	4.38	0.52	0.82	Very Good
6. Accessibility	4.13	0.64	0.75	Good
7. Reusability	4.25	0.46	0.78	Very Good
8. Standards Compliance	4.50	0.38	0.87	Excellent
9. Content Quality & Accuracy	4.38	0.52	0.82	Very Good
Overall Mean	4.38	0.47	0.83	Very Good

Rather than restating the tabulated scores, the validation pattern is interpreted as follows. The Motivation dimension attracted the highest rating (4.63), which is pedagogically meaningful: experts judged that the gamification layer operationalizes Self-Determination Theory coherently, pairing autonomy-supportive mechanics (level-choice, branching narratives) with competence scaffolds (ZPD-calibrated difficulty) and relatedness features (asynchronous

leaderboards, collaborative challenges). This triadic alignment, not gamification per se, is what reviewers credited, suggesting that needs-theory-grounded motivational design is perceived as the framework’s strongest pedagogical lever.

Accessibility scored lowest (4.13), reflecting concerns about learners with limited reading fluency, which is common in 3T schools with uneven baseline literacy. Two revision cycles introduced enlarged touch targets (e”48dp),

WCAG 2.1 AA contrast, audio narration of all text, and a visual-symbolic pathway for mathematics notation, strengthening alignment with UDL (CAST, 2018). Situating this within the international literature, our experience supports studies reporting that UDL implementation in low-resource settings is constrained less by principle than by infrastructure and material readiness (Kalogiannakis et al., 2021; Zourmpakis et al., 2023): the audio-first, low-text adaptations we adopted parallel accommodations recommended for under-resourced contexts and indicate that UDL is feasible offline when text dependence is minimized extending, rather than contradicting, prior UDL

evidence from better-resourced environments. Collectively, the four layers address the introduction’s gaps: they replace fragmented single-approach interventions (Salsabila et al., 2022; Maulidiansy et al., 2023) with a unified system, design explicitly for 3T constraints (Kalogiannakis et al., 2021; Zourmpakis et al., 2023), integrate gamification and adaptivity within one motivational–adaptive architecture (Chen et al., 2023; David & Weinstein, 2023), and establish computational thinking as a co-equal outcome in line with PISA 2025’s “Learning in the Digital World” domain. Table 7 positions ALF-STEAM-G against representative prior work.

Table 7. Comparative positioning of ALF-STEAM-G against prior work

Study / Framework	STEAM	Gamif.	Adaptive (ZPD/IRT)	Comp. Thinking	Numeracy	Offline/low-resource	Context
Salsabila et al. (2022)	X	X	X	X	✓	●	Interactive multimedia, numeracy (JHS, Indonesia)
Puspitasari (2022); Maulidiansy (2023)	✓	X	X	X	●	X	STEAM media for young learners (PAUD)
Kalogiannakis et al. (2021)	X	✓	X	X	X	X	Gamification in science (review)
Chen et al. (2023)	X	✓	X	X	X	X	Gamified e-books, language (flipped)
David & Weinstein (2023)	X	✓	X	X	X	X	Gamified experiential learning (SDT)
Boussakuk et al. (2021)	X	X	✓	X	X	X	Server-based IRT adaptive testing (CAT)
Zhang & Huang (2024)	X	✓	✓	X	X	X	Adaptive gamified assessment, blended
ALF-STEAM-G (this study)	✓	✓	✓	✓	✓	✓	Integrated offline system, Indonesia 3T

Note. –’ = feature fully addressed; Ñ% = feature partially addressed; ‘ = feature not addressed.

Practicality, Effectiveness Scope, and Roadmap to Efficacy Testing

In line with the editor’s requested R&D reporting structure, we distinguish three evaluative layers (design validity, practicality, and effectiveness) and state explicitly which are evidenced here. Design validity is supported by the LORI/Aiken’s V/CVR results above (RQ4). Practicality is supported at the design level by

expert usability ratings (Interaction Usability = 4.38) and by the offline, low-spec, teacher-accessible adaptations; however, large-scale student-response practicality data and classroom usability trials are scheduled for the product-development phase. Effectiveness on learning outcomes (pre-test/post-test gains, N-gain, and effect sizes) is deliberately not claimed in this paper, because no full product has yet been

deployed with students. The next phase will employ a randomized controlled (or quasi-experimental pre-post) design, reporting individual score trajectories (e.g., scatter plots of pre- vs post-scores) alongside aggregate N-gain and Cohen's d. Framing effectiveness as planned, pre-registered work avoids overclaiming while giving reviewers a concrete efficacy roadmap.

■ LIMITATIONS

Several limitations should be acknowledged. First, the study was conducted in a single regency (Pinrang, South Sulawesi); although purposive sampling captured urban, peri-urban, and 3T variation, generalization to other regions should be approached with caution. Second, the study validates the framework design through expert review (LORI, Aiken's V, CVR, inter-rater agreement) but does not yet provide efficacy evidence; consequently, it makes no causal claims about student learning gains. Third, the diagnostic instrument was calibrated with a 3PL model on 300 students, which is modest for stable estimation of the guessing parameter; item parameters should be re-estimated with a larger sample, and IRT-based marginal reliability should complement coefficient alpha. Fourth, several baseline indicators (teacher readiness, student motivation) rely on self-report and are subject to social-desirability bias. Fifth, economic estimates depend on projected five-year assumptions; the figures are indicative and warrant the sensitivity analysis reported in Appendix B as real implementation data accrue. Sixth, the 2GB-RAM offline performance is a design specification not yet benchmarked on physical devices; controlled performance testing is required before making runtime claims. Finally, inter-rater agreement was summarized with Fleiss' Kappa and complemented by ICC and Kendall's W; future validation should expand the expert panel and report all three indices.

■ CONCLUSION

This study constructed and, through expert design validation, established what is, to our knowledge, among the first adaptive learning frameworks to integrate STEAM pedagogy, gamification mechanics, ZPD-IRT personalization, and computational thinking assessment into a theoretically coherent, contextually calibrated design for elementary numeracy education in Indonesia's 3T regions. The needs analysis identified pedagogical stagnation, motivational disengagement, severe ICT inequities, and teacher-capacity constraints across 10 schools in Pinrang, which directly informed the architecture. Baseline data showed numeracy competency below national thresholds across all five domains (composite 50.1/100), with geometry (34.7) and computational thinking (48.3) most critical. Feasibility analysis confirmed pedagogical, technical, and economic viability (BCR = 1.42; NPV positive; IRR = 18.3%). The framework achieved a LORI mean of 4.38/5, Fleiss' Kappa of 0.74, and Aiken's V of 0.79.

Within the scope of design validation, the study contributes in three ways. Theoretically, ALF-STEAM-G coherently integrates constructivism, self-determination theory, ZPD-IRT adaptivity, UDL, and computational thinking for elementary numeracy in 3T areas. Methodologically, the DBR-integrated mixed-methods approach shows how systematic needs evidence can be translated into a defensible framework design. Practically, the low-computation ZPD-IRT design and offline-capable architecture are intended to address the infrastructure incompatibility that has hindered adaptive learning in 3T schools, a proposition to be confirmed empirically rather than asserted here. Accordingly, claims are bounded to the theoretical design and expert-validation evidence available, and definitive replicability and infrastructure-compatibility conclusions await field testing. Future research should advance to full

multimedia product development and a randomized controlled trial assessing impact on numeracy, computational thinking, and self-regulated learning, generating the causal efficacy evidence needed for national policy consideration.

■ REFERENCES

- Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142. <https://doi.org/10.1177/0013164485451012>
- Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational Researcher*, 41(1), 16–25. <https://doi.org/10.3102/0013189X11428813>
- Angrist, N., Aurino, E., Patrinos, H. A., Psacharopoulos, G., Vegas, E., Nordjo, R. E., & Wong, B. (2023). Improving learning in low- and lower-middle-income countries. *Journal of Benefit-Cost Analysis*, 14, 55–80. <https://doi.org/10.1017/bca.2023.26>
- Boussakuk, M., Bouchboua, A., Ghazi, M. E., Bekkali, M. E., & Fattah, M. (2021). Design of a computerized adaptive testing module for our dynamic adaptive hypermedia system. *International Journal of Emerging Technologies in Learning (iJET)*, 16(18), 113–130. <https://doi.org/10.3991/ijet.v16i18.23841>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Budiyono, B., Wiryanto, W., Suprayitno, S., & Primaniarta, M. G. (2023). *Persepsi mahasiswa pendidikan dasar terhadap gamifikasi dalam pendidikan STEAM* [Elementary education students' perceptions of gamification in STEAM education]. *Jurnal Obsesi: Jurnal Pendidikan Anak Usia Dini*, 7(3), 3591–3605. <https://doi.org/10.31004/obsesi.v7i3.4909>
- CAST. (2018). Universal Design for Learning Guidelines (Version 2.2). <http://udlguidelines.cast.org>
- Chen, C., Jamiat, N., & Mao, Y. (2023). The study on the effects of gamified interactive ebooks on students' learning achievements and motivation in a Chinese character learning flipped classroom. *Frontiers in Psychology*, 14, Article 1236297. <https://doi.org/10.3389/fpsyg.2023.1236297>
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research (3rd ed.)*. SAGE Publications.
- Dagiene, V., & Futschek, G. (2008). Bebras international contest on informatics and computer literacy: Criteria for good tasks. In R. T. Mittermeir & M. M. Syslo (Eds.), *Informatics in schools: Supporting computational thinking* (pp. 19–30). Springer.
- Dahlia, D., Oci, N., Pujiastuti, H., & Mutaqin, A. (2024). *Peran soft skill dan hard skill dalam kemampuan numerasi siswa: Systematic literature review* [The role of soft skills and hard skills in students' numeracy abilities: Systematic literature review]. *EDUKATIF: Jurnal Ilmu Pendidikan*, 6(4), 4185–4198. <https://doi.org/10.31004/edukatif.v6i4.7471>
- David, L., & Weinstein, N. (2023). A gamified experiential learning intervention for engaging students through satisfying needs. *Journal of Educational Technology Systems*, 52(1), 52–78. <https://doi.org/10.1177/00472395231174614>
- Heryani, Y., Kartono, K., Wijayanti, K., & Dewi, N. R. (2023). Analysis of students' mathematical literacy ability in solving

- HOTS problems in minimum competency assessment. *Journal of Higher Education Theory and Practice*, 23(16). <https://doi.org/10.33423/jhetp.v23i16.6470>
- Jusniani, N., Nursofa, W., & Rahmi, D. F. (2023). *Pengaruh kemampuan pemecahan masalah matematik dan motivasi belajar terhadap numerasi siswa kelas 4 SD* [The influence of mathematical problem-solving abilities and learning motivation on the numeracy of grade 4 elementary school students]. *Journal of Contemporary Issue in Elementary Education*, 1(2), 72–85. <https://doi.org/10.33830/jciee.v1i2.6469>
- Kabilan, M. K., Annamalai, N., & Chuah, K. M. (2023). Practices, purposes and challenges in integrating gamification using technology: A mixed-methods study on university academics. *Education and Information Technologies*, 28(11), 14249–14279. <https://doi.org/10.1007/s10639-023-11723-7>
- Kalogiannakis, M., Papadakis, S., & Zourmpakis, A.-I. (2021). Gamification in science education: A systematic review of the literature. *Education Sciences*, 11(1), Article 22. <https://doi.org/10.3390/educsci11010022>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.
- Maulidiansy, H., Novaliyosi, & Hendrayana, A. (2023). *E-Motivector (e-modul interaktif berbantuan video creator) yang mendukung pembelajaran STEAM* [E-Motivector (an interactive e-module assisted by a video creator) that supports STEAM learning]. *Proximal: Jurnal Penelitian Matematika dan Pendidikan Matematika*, 6(2), 336–350. <https://doi.org/10.30605/proximal.v6i2.2742>
- Masraria, Pajariantanto, H., Masnur, & Ching Yang, C. (2026). The mediating role of student feedback in math playground game learning: Effects on perceived ease, motivation, and learning outcomes. *Jurnal Kependidikan*, 12(1), 177–188. <https://doi.org/10.33394/jk.v12i1.19625>
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook (2nd ed.)*. SAGE Publications.
- Nesbit, J., Belfer, K., & Leacock, T. (2009). *Learning object review instrument (LORI) user manual (Version 2.0)*. E-Learning Research and Assessment Network.
- Nurhasnah, N., Sepriyanti, N., & Kustati, M. (2024). Learning theories according to constructivism theory. *Journal International Inspire Education Technology*, 3(1), 19–30. <https://doi.org/10.55849/jiiet.v3i1.577>
- Nurhayati, N., Asrin, A., & Dewi, N. K. (2022). *Analisis kemampuan numerasi siswa kelas tinggi dalam penyelesaian soal pada materi geometri di SDN 1 Teniga* [Analysis of the numeracy abilities of high-grade students in solving problems on geometry material at SDN 1 Teniga]. *Jurnal Ilmiah Profesi Pendidikan*, 7(2b), 723–730. <https://doi.org/10.29303/jipp.v7i2b.678>
- Puspitasari, E., Solfiah, Y., & Zulkifli, N. (2022). *Pengembangan scanbook untuk pembelajaran berbasis STEAM di lembaga PAUD* [Development of scanbooks for STEAM-based learning in PAUD institutions]. *Jurnal Obsesi: Jurnal*

- Pendidikan Anak Usia Dini, 6(6), 6173–6188. <https://doi.org/10.31004/obsesi.v6i6.2025>
- Putrawangsa, S., & Hasanah, U. (2022). *Analisis capaian siswa Indonesia pada PISA dan urgensi kurikulum berorientasi literasi dan numerasi* [Analysis of Indonesian students' achievements in PISA and the urgency of a literacy and numeracy-oriented curriculum]. *EDUPEDIKA: Jurnal Studi Pendidikan dan Pembelajaran*, 1(1), 1–14. <https://doi.org/10.60004/edupedika.v1i1.1>
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78. <https://doi.org/10.1037/0003-066X.55.1.68>
- Salsabila, N. H., Baidowi, B., Azmi, S., & Lu'luilmaknun, U. (2022). Interactive multimedia oriented to the mathematical numeracy skills of junior high school students. *Al Khawarizmi: Jurnal Pendidikan dan Pembelajaran Matematika*, 6(2), 130–142. <https://doi.org/10.22373/jppm.v6i2.15118>
- Saleha, S., Nadar, N., & Masnur, M. (2026). Developing a TPACK-based science module through project-based learning to enhance elementary students' 4Cs skills. *Jurnal Prima Edukasia*, 14(1), 100–112. <https://doi.org/10.21831/jpe.v14i1.89864>
- Silvia, R., & Asdarina, O. (2024). *Kemampuan numerasi siswa pada materi operasi pecahan dengan implementasi model problem based learning (PBL)* [Students' numeracy skills on fraction operations material with the implementation of the problem based learning (PBL) model]. *EMTEKA: Jurnal Pendidikan Matematika*, 5(1), 62–74. <https://doi.org/10.24127/emteka.v5i1.5516>
- Susandi, A. D., Kandaga, T., Basri, H., & Rindastri, V. (2025). Mapping elementary students' numeracy literacy skills: A comprehensive analysis based on six indicators. *Smart Society Community Service and Empowerment Journal*, 5(2), 259–270. <https://doi.org/10.58524/martsociety.v5i2.852>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Waluyo, R., Kusuma, L. L., & Zamora, H. P. (2024). *Pemanfaatan teknologi untuk meningkatkan numerasi peserta didik SMP N 1 Kertanegara* [Utilization of technology to improve numeracy of students at SMP N 1 Kertanegara]. *Jurnal Pengabdian Masyarakat (Abdira)*, 4(3), 9–17. <https://doi.org/10.31004/abdira.v4i3.491>
- Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3), 33–35. <https://doi.org/10.1145/1118178.1118215>
- Zhang, Z., & Huang, X. (2024). Exploring the impact of the adaptive gamified assessment on learners in blended learning. *Education and Information Technologies*, 29(14), 21869–21894. <https://doi.org/10.1007/s10639-024-12708-w>
- Zourmpakis, A.-I., Kalogiannakis, M., & Papadakis, S. (2023). Adaptive gamification in science education: An analysis of the impact of implementation and adapted game elements on students' motivation. *Computers*, 12(7), 143. <https://doi.org/10.3390/computers12070143>