AI-Powered Computer-Based Assessment for Literacy and Numeracy: Multi-School Validation in Indonesian Secondary Education

Anna Fitri Hindriana[1,*], Rio Priantama[2], Ina Setiawati[3], and Alin Rizki Pratami[3]

[1]Program Studi Pendidikan Biologi, Sekolah Pascasarjana, Universitas Kuningan, Kuningan, Indonesia
[2]Program Studi Teknik Informatika, Fakultas Komputer, Universitas Kuningan, Kuningan, Indonesia
[3]Program Studi Pendidikan Biologi, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Kuningan, Kuningan, Indonesia
Email: anna@uniku.ac.id (A.F.H.); rio.priantama@uniku.ac.id (R.P.); ina.setiawati@uniku.ac.id (I.S.); alinpratami@gmail.com (A.R.P.)
*Corresponding author

Manuscript received September 1, 2025; revised October 15, 2025; accepted December 5, 2025; published May 15, 2026

***Abstract*—Accurately diagnosing literacy and numeracy competencies remains a persistent challenge in developing countries despite significant investments in educational technology. This study developed and validated Asesmen Kompetensi Minimum (AKM) Online, an Artificial Intelligence (AI)-enhanced Computer-Based Assessment (CBA) system that integrates Fisher–Yates Shuffle and Regular Expression algorithms to enable secure test randomization, automated response validation, and real-time diagnostic feedback at scale. Through multi-school implementation across three contrasting socioeconomic contexts (N = 552), the system generated 17,049 student responses and demonstrated high technical reliability (AI scoring accuracy = 94.35%). Results revealed severe and consistent numeracy deficiencies (95.8–100% requiring intervention) and substantial variation in literacy needs (53.9–78.46%), indicating that contextual factors influence literacy more strongly than numeracy. Mathematical reasoning showed complete failure (0% proficiency), a finding corroborated through expert validation, cross-context replication, and alignment with national and international benchmarks). Expert evaluations confirmed high system quality in both educational assessment (81%) and informatics performance (82%). The study provides empirical evidence that AI-supported CBA can overcome major limitations of existing national assessments by enabling whole-population testing with immediate, actionable competence profiles. The scalability, transparency, and resource-efficient design of AKM Online underscore its potential for broader adoption in developing countries seeking reliable, evidence-based approaches to foundational literacy and numeracy assessment.**

***Keywords*—computer-based assessment, Artificial Intelligence (AI), literacy, numeracy, educational technology**

I. Introduction

Despite major efforts to expand the use of digital technologies in education, many countries continue to face substantial challenges in producing accurate, timely, and actionable assessments of students' literacy and numeracy skills. Persistent learning gaps—particularly in developing contexts where instructional resources and assessment capacity are limited—underscore the need for more responsive and diagnostically precise measurement systems. Recent global reports reveal that a significant proportion of children in low- and middle-income countries fail to achieve basic proficiency in foundational competencies, signaling long-standing systemic issues that conventional assessment approaches have been unable to resolve [1–3]. These conditions highlight the urgency of assessment innovations capable of providing reliable individual-level insights to support targeted learning interventions [4, 5].

Computer-Based Assessment (CBA) with algorithmic integration offers innovative solutions through assessment personalization, response pattern analysis, and real-time feedback supporting differentiated learning [6, 7]. Recent meta-analyses demonstrate that well-designed CBA systems can improve learning outcomes with effect sizes ranging from moderate to large (Cohen's d = 0.62–1.18) compared to traditional paper-based assessments [8, 9].

Indonesia illustrates these systemic challenges. Results from the 2018 Programme for International Student Assessment (PISA) ranked the country 73rd of 79 participating nations in mathematics, with 70% of students falling below the baseline proficiency level in problem-solving, far below the OECD average [10, 11]. The national Minimum Competency Assessment (Asesmen Kompetensi Minimum/AKM), introduced in 2021, represents a major policy effort toward improving foundational competencies. However, its sampling approach—testing only 45 students per school—and significant reporting delays limit its diagnostic usefulness for classroom-level instructional decision-making [12, 13].

The term "AI-enhanced" in this study refers to intelligent automation for pattern recognition, adaptive randomization, and automated student classification. We use a rule-based approach rather than machine learning to ensure system transparency, minimize computational requirements, and enable deployment across diverse school infrastructures. Our study addresses gaps in the National Minimum Competency Assessment (AKM), which tests only 45 students per school and provides delayed results without individual breakdowns across literacy and numeracy indicators. AKM Online enables full population assessment with immediate individual diagnostic profiles across all competency indicators. This represents a significant advancement in assessment approach and scalability. While advanced methods like Natural Language Processing (NLP)-based scoring or Machine Learning (ML)-driven adaptive testing could enhance the system, our rule-based implementation proves practically effective in developing country contexts where infrastructure and ML expertise are limited.

The "AI-enhanced" designation in this research refers to intelligent automation capabilities specifically real-time pattern recognition, adaptive randomization, and automated decision-making for student classification implemented through rule-based algorithms (Fisher-Yates Shuffle and

Regular Expression) rather than machine learning approaches. This pragmatic approach ensures system transparency, minimal computational requirements, and deployment feasibility across diverse school infrastructures while maintaining high technical reliability [14]. Our system addresses fundamental gaps in Indonesia's National Assessment by enabling whole-population assessment with immediate individual-level diagnostic profiles across all six competency indicators: literacy indicators (Information Finding, Interpretation-Integration, Evaluation-Reflection) and numeracy indicators (Understanding, Application, Reasoning). While advanced AI approaches such as NLP-based automated scoring or machine learning-driven adaptive testing would further enhance capabilities, our rule-based implementation demonstrates practical effectiveness in resource-constrained developing country contexts where infrastructure and expertise for sophisticated ML implementations may be limited [15, 16].

Systematic reviews reveal significant literature gaps [5, 17]. Most CBA studies focus on technical validation or small-scale implementations with samples under 200 participants [18, 19], with limited empirical validation across diverse socioeconomic contexts [20, 21]. He *et al.* [22] demonstrating system effectiveness under real conditions with large samples and diverse geographical contexts remains limited. Furthermore, algorithmic integration with empirical validation of educational outcomes in multi-school settings with different socioeconomic characteristics represents an underexplored area [23], particularly critical for developing countries where infrastructure disparities create unique implementation challenges [24, 25]. Current systems frequently provide numerical scores without actionable insights for improvement, lacking mechanisms to diagnose specific competency deficiencies or recommend targeted interventions [26, 27].

The Fisher-Yates algorithm ensures assessment integrity through mathematically proven uniform distribution properties, preventing pattern memorization while maintaining fairness [16]. Regular Expression algorithms enable sophisticated automated answer validation while maintaining high scoring accuracy through flexible pattern recognition that accommodates response format variations [25]. The system is designed with cloud-based architecture accessible through various devices (desktop computers, laptops, tablets, smartphones) to address infrastructure disparities common in developing countries [28], while providing automatic student classification into four competency levels and real-time feedback to support immediate instructional adjustments [29].

The present study addresses these gaps by developing and validating an AI-enhanced CBA system—AKM Online—designed specifically to support full-population diagnostic assessment of literacy and numeracy in Indonesian secondary education. The system incorporates core algorithms to ensure assessment integrity, automated classification, and real-time feedback, and is implemented across multiple schools representing diverse socioeconomic contexts. The research pursues four objectives: (1) developing a CBA system that meets principles of validity, reliability, and practical usefulness; (2) evaluating system performance through multi-context implementation with 552 students; (3) analyzing literacy and numeracy competency patterns to inform evidence-based educational policy; and (4) demonstrating the system's scalability and transferability for use in developing country contexts.

The main contributions of this research include: (1) empirical validation of an AI-enhanced CBA system through large-scale implementation across multiple contexts, addressing the gap in ecological validity research; (2) demonstration of comprehensive competency diagnosis capabilities with automatic classification and real-time feedback in resource-constrained settings; (3) identification of consistent numeracy crisis patterns across diverse contexts, providing robust evidence for urgent policy intervention; and (4) methodological contributions through multi-level validation approaches applicable for educational technology research in developing countries. These findings are expected to provide a scientific foundation for assessment transformation in Indonesia and offer a transferable model for other developing countries facing similar challenges in literacy and numeracy education [30, 31].

II. Literature Review

A. Evolution of Computer-Based Assessment Technologies

The integration of digital technologies in educational assessment has reshaped how student competencies are measured, interpreted, and supported. Over the past decade, Computer-Based Assessment (CBA) has advanced significantly, enabling more efficient administration, enhanced scoring accuracy, and improved diagnostic capabilities compared to conventional paper-based assessments. Meta-analytic evidence across 127 studies demonstrates that well-designed CBA systems yield moderate to large improvements in learning outcomes (Cohen's d = 0.62–1.18), highlighting their pedagogical advantages [32, 33].

Computerized Adaptive Testing (CAT), one of the most sophisticated CBA developments, has shown particularly strong performance in mathematics assessment (weighted effect size d = 0.89, 95% CI: 0.74–1.04), offering real-time difficulty adjustment and individualized precision [34]. Additionally, emerging CBA models leverage log-based analytics to examine response patterns, cognitive processes, and student engagement, extending assessment beyond final answers to capture more nuanced representations of learning behaviour [15, 35].

Recent innovations include response-time analytics capable of estimating cognitive load with over 94% accuracy and machine-learning models that detect disengaged responding with accuracy rates between 91% and 96% [16]. Despite these advances, systematic reviews indicate that approximately 73% of digital assessments still function as digitized replicas of paper-based tests—failing to exploit technological affordances such as adaptivity, automation, or real-time feedback [17, 19]. This gap signals the need for more transformative CBA designs capable of addressing structural limitations in traditional assessment systems.

B. Artificial Intelligence Algorithms in Educational Assessment

The integration of specific AI algorithms for educational

assessment represents a critical frontier in contemporary research. Fisher-Yates Shuffle algorithms have emerged as the gold standard for question randomization, ensuring assessment integrity and preventing pattern memorization through mathematically proven uniform distribution properties [24]. Parallel developments in Regular Expression (RegEx) algorithms enable sophisticated automated answer validation while maintaining scoring accuracy through flexible pattern recognition that accommodates response format variations [25].

Recent validation studies demonstrate that Fisher-Yates implementations achieve randomization quality metrics exceeding 95% uniformity distribution, with Chi-square goodness-of-fit tests consistently confirming statistical randomness [28]. Contemporary Regex-based validation systems report precision rates above 94% with recall consistency maintaining optimal F1-Scores for automated scoring reliability [36]. These technical advances address critical limitations in earlier CBA systems where inadequate randomization and rigid answer matching compromised assessment validity.

C. Literacy and Numeracy Assessment in Developing Countries Context

International assessments reveal persistent challenges in literacy and numeracy competency measurement, particularly in developing countries where 70% of students fail to meet basic proficiency standards [37]. Indonesia's PISA 2018 mathematics performance (379/490, ranking 73rd/79 countries) exemplifies this crisis, with 70% of students unable to achieve Level 2 proficiency [12, 14].

These results underscore long-standing tensions between formal mathematical instruction and practical numeracy application, reflecting gaps in conceptual understanding and reasoning skills [38, 39].

Recent studies using machine-learning models further show that affective factors, such as mathematics anxiety, significantly predict performance in both literacy and numeracy. These psychological barriers are particularly pronounced in learning environments where cognitive and emotional supports are limited [40]. Together, these findings emphasize the need for assessment systems that can detect differentiated competency profiles and provide actionable insights for targeted intervention.

D. Emerging Trends in Adaptive Learning Analytics

The convergence of learning analytics and adaptive assessment represents the current research frontier. Contemporary systems demonstrate real-time competency classification and personalized intervention recommendations through comprehensive response pattern analysis [41, 42]. Recent advances enable analysis of not only final answers but also problem-solving processes, response times, and strategic approaches, providing multidimensional competency profiles [43].

However, scalability of these features in resource-constrained environments remains a critical research question, with limited empirical evidence demonstrating effectiveness across diverse infrastructures and varying digital literacy levels among stakeholders.

III. Materials and Methods

A. Research Design and Sample

This study employed a mixed-methods, multi-phase validation design to assess the effectiveness of the AKM Online Application across diverse educational contexts. The research incorporated three sequential phases: expert validation, multi-school pilot implementation, and comprehensive empirical analysis with multilevel statistical modelling. The design enabled both technical validation of AI algorithms and educational effectiveness assessment of competency identification.

Furthermore, the study involved 552 Grade 8 students selected through purposive sampling across three junior secondary schools in West Java and Central Java provinces. Grade 8 was chosen because this level aligns with Indonesia's national Minimum Competency Assessment expectations and provides sufficient time for remedial intervention prior to upper secondary school.

School selection criteria included: computer laboratory availability (minimum 20 workstations), stable internet connectivity (10–50 Mbps bandwidth), diverse socioeconomic contexts, and administrative support for research participation.

Participant distribution:

- School A (Central Java, rural; low-income): 26 students
- School B (West Java, semi-urban; middle-income): 31 students
- School C (West Java, urban; mixed-income): 495 students across seven classes

Additionally, a detailed analysis focused on School C (N = 495) was conducted due to adequate sample size for robust statistical testing and multilevel modeling. Class distribution comprised: 8A (34 students), 8B (35 students), 8C (34 students), 8D (42 students), 8E (32 students), 8F (39 students), and 8G (39 students). This implementation generated 17,049 individual assessment responses across all competency indicators.

B. System Design and Technical Specifications

The AKM Online Application was developed by the research team at Universitas Kuningan specifically for this study, integrating evidence-based assessment principles with algorithmic automation for scoring and adaptive question delivery. The system represents original research contribution rather than commercial software or Ministry of Education collaboration, designed specifically to address research objectives while maintaining potential for broader implementation.

The system operates as a web-based platform accessible via modern web browsers on desktop computers, laptops, tablets, and smartphones. This architecture ensures cross-device compatibility while maintaining assessment integrity through centralized server-side processing and real-time data synchronization. All assessment logic, randomization algorithms, scoring procedures, and data storage are managed server-side, preventing client-side manipulation while enabling consistent user experience across devices with varying specifications.

In relation to the network and hardware requirements, the AKM online application required stable internet connectivity

with bandwidth ranging 10–50 Mbps across participating schools, accommodating variability in infrastructure quality across rural, semi-urban, and urban contexts. Each computer laboratory maintained minimum technical standards: individual workstations with modern web browsers), stable network connection, and standardized screen resolution for consistent user experience. No specialized hardware or software installation was required standard school computer laboratories with internet access were sufficient.

Equally important are the core system features and functionality which makes the system is able to provide a comprehensive assessment management capabilities through integrated modules:

- **Administration and Access Control.** Role-based authentication (Teacher, Student, Administrator) with Back-end API (Express.js) and Front-end spec enables: (1) students to access assigned assessments, personal dashboard, and historical performance; (2) teachers to manage class rosters, monitor student performance, and generate detailed reports; (3) administrators to oversee configuration, user management, and system monitoring with multi-layer security ensuring data privacy.
- **Dynamic Question Presentation.** Assessments feature randomized instruction-to-exam distributions with clear question stems, appropriate formatting, and Fisher-Yates randomized multiple-choice options. The system displays countdown timers, provides preview of future items, and allows navigation to previous items before final submission.
- **Built-in Countdown Timer.** Visible countdown automatically submits responses when time expires, ensuring fairness and standardization. Students receive warnings at 15, 10, and 5 min remaining, enabling time management and preventing accidental time loss.
- **Secure Answer Submission.** Upon completion or time expiration, responses are submitted via encrypted HTTPS with integrity metadata. Students receive confirmation messages with unique submission ID for record-keeping.
- **Real-time Result Generation and Reporting.** Results are calculated immediately using Regular Expression algorithm. Students instantly view overall competency classifications (Need-Special Intervention, Basic, Proficient, Advanced), detailed breakdown across six indicators with percentage accuracy, competency-specific feedback identifying strengths and weaknesses, and actionable recommendations prioritized by deficiency severity. Teachers access aggregated class reports for instructional planning.
- **Longitudinal Tracking and Analytics.** The system maintains complete assessment history with date, score, competency profiles, progress tracking showing improvement trends, comparative analysis across multiple administrations, and detailed item-level response patterns for diagnostic purposes.

Besides the comprehensive assessment management, the assessment integrity, including cheating prevention is maintained through multiple complementary security mechanisms:

- **Question Randomization:** Each student receives questions in individually randomized sequence generated using Fisher-Yates algorithm ensuring uniform probability distribution across all possible orderings. This prevents students from copying answers based on question position, as "Question 1" differs for each student. Randomization occurs server-side upon assessment initiation, maintaining consistent sequence for individual student if they pause and resume assessment.
- **Answer Option Randomization:** Within each multiple-choice question, answer options (A, B, C, D) are randomly shuffled for each student, further preventing position-based copying. Correct answer may appear in any position, ensuring students must engage with content rather than memorizing answer patterns.
- **Timing Control and Navigation Monitoring:** The system records comprehensive interaction timestamps, routing, behavioral audit trail, login time and IP address, question access sequence and timestamps, and time spent on each item with flowchart recording every answer change (initial responses and modifications), forms enabled for review, pause/resume events, and submission timestamp. These logs enable post-assessment behavior analysis for anomaly detection including: (a) excessively rapid submissions (<30 s per question), (b) repeating exact identical response timing patterns between students suggesting collusion, and (c) suspicious navigation patterns.

Lastly, to insure the data and quality of the assessment result, the system operates on cloud-based infrastructure (Amazon Web Services EC2 instances) providing high availability through redundant servers and automatic failover preventing disruption from individual server failures, scalability enabling concurrent access by hundreds of students without performance degradation, and geographic distribution reducing latency for users across Indonesian archipelago. Quality assurance protocols include: automatic connectivity monitoring with real-time alerts for network issues, response validation ensuring all data submitted by students is properly received and stored, data integrity verification through checksums confirming transmitted data matches received data, automated backup procedures creating redundant copies of all assessment data at 15 min intervals, and comprehensive error logging capturing any system anomalies for troubleshooting.

Technical support team monitored all assessment sessions remotely, ready to resolve issues immediately. Success rate exceeded 99.5% (all but 8 of 552 students completed assessments without technical interruptions), confirming system reliability under real-world implementation conditions.

C. Assessment Framework

The assessment instrument was developed by the Center for Educational Assessment, Ministry of Education and Culture, Republic of Indonesia, ensuring alignment with national Minimum Competency Assessment (AKM) standards. Item development followed rigorous psychometric procedures including expert review, pilot testing, and validation across diverse student populations nationally. This national validation ensures that items reflect curriculum standards and are appropriate for Grade 8 students across Indonesian educational contexts.

The assessment comprised 65 items distributed across six competency indicators following cognitive complexity

hierarchies. Table 1 presents the comprehensive test blueprint:

Table 1. Test blueprint and item distribution

Competency Domain	Indicator Code	Indicator Name	Cognitive Level	Item Count	Percentage	Total Items
Literacy	INF-1	Information Finding	Access and Retrieve	13	40%	**31**
Literacy	IPR-2	Interpretation and Integration	Interpret and Integrate	12	40%	**31**
Literacy	EVA-3	Evaluation and Reflection	Evaluate and Reflect	6	20%	**31**
Numeracy	PHM-1	Understanding	Knowing	14	40%	**34**
Numeracy	TRP-2	Application	Applying	13	40%	**34**
Numeracy	NLR-3	Reasoning	Reasoning	7	20%	**34**

Note: Item distribution reflects national AKM framework priorities, with 40% allocation to foundational skills, 40% to application/integration, and 20% to higher-order thinking skills

The indicators are categorised into literacy and numeracy domain. The literacy domain INF-1 (Information Finding—Access and Retrieve): Measures ability to locate and access explicitly stated information within texts. The 40% allocation (13 items) ensures robust measurement of foundational reading comprehension, providing a reliable assessment of basic literacy skills essential for more advanced competencies.

Meanwhile, the IPR-2 (Interpretation and Integration): Assesses capacity to understand implicit information, make inferences, and integrate information across text segments. This indicator represents intermediate-level literacy requiring active meaning construction beyond surface-level comprehension. Lastly, the EVA-3 (Evaluation and Reflection): Evaluates higher-order critical reading skills including assessing source credibility, analyzing author's purpose, evaluating argument quality, and connecting text content to external knowledge and personal experience. The 20% allocation (6 items) reflects the advanced nature of these skills, which represent the pinnacle of literacy development requiring metalinguistic awareness and critical thinking.

Likewise, the numeracy competency indicators also covers three codes: PHM-1, TRP-2, and NLR-3. PHM-1 (Understanding—Knowing) measures comprehension of fundamental mathematical facts, concepts, procedures, and tools. The 40% allocation (14 items) ensures thorough assessment of foundational numeracy essential for application and reasoning. Then, the TRP-2 (Application—Applying) assesses ability to apply mathematical concepts and procedures in routine, structured situations. Items require students to execute standard algorithms, apply formulas in straightforward problems, solve routine word problems, and use mathematical tools appropriately in familiar contexts. This represents practical numeracy skills applicable to everyday mathematical tasks.

Lastly, the NLR-3 (Reasoning) which evaluates capacity for mathematical reasoning in non-routine, complex problem-solving situations. The 20% allocation (7 items) reflects advanced cognitive demands of mathematical reasoning requiring flexible thinking and creative problem-solving approaches.

These indicators are prior subjected to validity tests covers (1) Content validity through Ministry expert development ensuring alignment with curriculum standards and educational objectives, (2) Structural validity through hierarchical competency structure aligned with established learning progressions and cognitive development theories, (3) Substantive validity through sufficient items per indicator (6–14 items) enabling reliable measurement while avoiding excessive testing burden, and (4) External validity through consistency with national AKM framework used across Indonesian educational system. The explicit connection between theoretical constructs (literacy and numeracy as defined in national standards) and operational measures (specific item types and distributions) strengthens evidence that the assessment validly measures intended competencies rather than irrelevant factors.

D. Artificial Intelligence Algorithm Implementation

The AI-enhanced designation refers to intelligent automation capabilities that ensure system transparency, minimal computational requirements, and deployment feasibility across diverse school infrastructures while maintaining high technical reliability. The system employs two core algorithms for ensuring assessment integrity and scoring accuracy.

The first algorithm is the Fisher-Yates shuffle that was employed for question randomisation to ensure assessment integrity and prevent pattern memorization. The randomisation quality was validated using Chi-square goodness-of-fit tests comparing observed versus expected uniform distribution, with acceptance criteria of $p > 0.05$.

The second algorithm is Regular Expression (Regex) which Automate the answer validation for flexible response matching while maintaining scoring accuracy. Performance validation employed confusion matrix analysis, calculating precision, recall, and F1-Score metrics.

E. Data Collection and Validation

The data collection process is divided into three steps; validation, implementation, and quality assurance phases. In the data validation phase, four independent experts (two educational assessment, two informatics) with doctoral degrees and 10+ years experience evaluated the application using structured 5-point Likert scales. Learning experts assessed competency alignment (85%), item construction clarity (80%), result stability (81%), stimulus appropriateness (75%), and bias absence (81%), averaging 81% (Excellent). Informatics experts assessed system reliability (82%), data security (85%), performance optimization (81%), UI/UX quality (83%), anti-cheating mechanisms (85%), and compatibility (75%), averaging 82% (Excellent). Inter-rater reliability: initial $\alpha K = 0.742–0.769$, post-consensus $\alpha K = 0.847–0.863$.

In the implementation phase, the assessment sessions were ensured to followed standardised protocols by with trained supervisiors maintaining consistent conditions across all schools. Each session included pre-assessment orientation, technical verification, 90 min assessment period, and immediate data backup verification. Real-time technical support ensured continuous system operation and immediate issue resolution.

Subsequently, the data quality control phase included

automatic range validation (0–100% accuracy scores), response completeness verification, timestamp accuracy checking, and statistical outlier identification. Post-collection validation confirmed data integrity, logical consistency, and missing data patterns before statistical analysis.

F. Statistical and Psychometric Analysis

The statistical analysis plan covers descriptive analysis and inferential statistical testing. The performance analysis calculated means, standard deviations, and 95% confidence intervals for all competency indicators and achievement levels. While the cross-tabulation examined competency distribution patterns across schools and classes. The coefficient of variation assessed performance consistency within and between educational contexts.

For the inferential testing, the One-way ANOVA tested performance differences across the three schools with post-hoc multiple comparison corrections using Tukey's HSD method. Also the Chi-square tests was conducted to examine the associations between schools and competency level distributions. The Effect sizes were calculated using eta-squared (η^2) with interpretation criteria: small ($\eta^2 = 0.01$), medium ($\eta^2 = 0.06$), and large ($\eta^2 = 0.14$). Additionaly, the hierarchical linear modeling was accounted for nested data structure (students within classes within schools) using school and class as random effects to assess between-group variations while controlling for clustering effects.

Another essential statistical procedure is the Literacy-Numeracy Comparison using Cohen's d to calculated the effect sizes for performance differences between literacy and numeracy competencies with magnitude interpretation: small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$). The performance indicator differences for the six indicators were measured using repeated ANOVA compared performance across the six competency indicators with Greenhouse-Geisser correction for sphericity violations when detected. For the class-level analysis, the Intraclass Correlation Coefficients (ICC) were considered to quantify variance attributable to class-level factors versus individual student differences. All tests employed $\alpha = 0.05$ with effect size reporting. Missing data (<2%) were handled through listwise deletion after confirming MCAR patterns.

Following this, the psychometric analysis was conducted on School C data ($N = 495$) to ensure adequate sample size for CFA and IRT modelling. School C was selected due to sufficient sample size, contextual homogeneity, and representation of the urban mixed-income context. Multi-school validation ($N = 552$) subsequently examined measurement consistency.

Confirmatory Factor Analysis tested the hypothesized two-factor structure (Literacy: INF-1, IPR-2, EVA-3; Numeracy: PHM-1, TRP-2, NLR-3) using maximum likelihood estimation with robust standard errors (MLR) to account for potential non-normality in score distributions.

Model fit was evaluated using multiple indices following established guidelines: Comparative Fit Index (CFI ≥ 0.90), Tucker-Lewis Index (TLI ≥ 0.90), Root Mean Square Error of Approximation (RMSEA ≤ 0.08), and Standardized Root Mean Square Residual (SRMR ≤ 0.08) [44]. Convergent validity was assessed through Average Variance Extracted (AVE > 0.50) and Composite Reliability (CR > 0.70); discriminant validity was confirmed when √AVE exceeded inter-construct correlations [45].

Two-Parameter Logistic IRT modeling estimated item parameters for each competency indicator using marginal maximum likelihood estimation. The 2PL model estimates: (1) discrimination parameter (a), representing the indicator's ability to differentiate between students of different ability levels (higher values indicate steeper item response curves and better discrimination), with $a > 1.0$ considered acceptable and $a > 1.7$ considered excellent, and (2) difficulty parameter (b), representing the ability level at which students have 50% probability of correct response, typically ranging from −3 to +3 logits on the standardized ability scale [46].

Model fit was assessed via −2 Log-Likelihood ratio tests comparing the 2PL model against more restrictive Rasch models (which assume equal discrimination) and more flexible 3PL models (which add pseudo-guessing parameters). Additionally, item-level fit was examined using S − χ^2 statistics ($p > 0.01$ acceptable), which compare observed versus expected response patterns [47]. Test Information Functions identified optimal measurement precision ranges. Internal consistency was evaluated using Cronbach's α, McDonald's ω, and IRT-based marginal reliability [48]. Multi-group CFA examined measurement invariance across all three schools (N = 552) testing configural, metric, and scalar invariance using $\Delta CFI \leq 0.010$ and $\Delta RMSEA \leq 0.015$ criteria [49]. Despite small sample sizes in Schools A ($n = 26$) and B ($n = 31$), multi-group analysis examined structural consistency across diverse contexts. All analyses were conducted using R 4.3.2 with specialised packages: lavaan for CFA and measurement invariance [50] mirt for IRT modeling [51], and semTools for additional psychometric indices [52].

In relation to the AI Algorithm, the Fisher-Yates effectiveness validated through Chi-square tests confirming consistent randomization quality. Regular Expression performance assessed using precision-recall analysis with confusion matrix evaluation comparing automated vs. expert scoring.

Lastly, to address concerns about item appropriateness given observed low performance, particularly the 0% accuracy in mathematical reasoning (NLR-3), four content experts independently validated assessment items using structured 5-point Likert scales. The expert panel comprised two mathematics education specialists and two curriculum specialists.

The literacy items across all three indicators (INF-1, IPR-2, EVA-3) achieved an average rating 4.3/5.0 (range: 4.0–4.6), confirming alignment with Grade 8 literacy competencies according to national curriculum standards (Kurikulum Merdeka) and developmental appropriateness for 13–14 year old students. Experts confirmed that reading passages reflected appropriate complexity, vocabulary suited grade level, and questions accurately assessed intended competencies without cultural or socioeconomic bias. Numeracy items, with particular attention to NLR-3 given the observed zero performance, achieved an average rating 4.2/5.0 (range: 4.0–4.5). Experts confirmed: (a) curriculum alignment items accurately reflect Grade 8 mathematics curriculum standards (Kurikulum Merdeka KD 3.3–3.4) including mathematical reasoning competencies specified in

national standards, (b) assessment specification match items align with National Assessment specifications for numeracy reasoning including problem-solving in novel contexts, pattern recognition, and mathematical justification, and (c) developmental appropriateness items represent cognitive demands appropriate for Grade 8 students who have received standard mathematics instruction through Grade 7, including foundational skills in arithmetic, basic algebra, geometry, and logical reasoning.

The Pilot testing (N = 45 students from a school not included in main implementation) confirmed item comprehensibility students understood instructions and could engage with items regardless of correctness and administration procedures including timing adequacy and technical functionality. This expert validation demonstrates that assessment design is appropriate for the grade level according to national curriculum standards and expert professional judgment. The observed low performance, including zero mathematical reasoning accuracy, reflects genuine student ability patterns rather than design flaws such as unclear wording, inappropriate difficulty beyond curriculum standards, or cultural bias.

G. Limitations

The study acknowledged geographic limitation to two provinces, though diverse socioeconomic contexts enhanced external validity. Cross-sectional design provided snapshot assessment effectiveness. Technical infrastructure requirements limited implementation to schools with adequate facilities. Our rule-based implementation prioritizes sustainability and broad accessibility over algorithmic sophistication, appropriate for developing country contexts where ML infrastructure may be limited. Sample size adequacy confirmed through post-hoc power analysis (>0.80 for medium-large effect sizes). Future enhancements could incorporate NLP-based scoring for open-ended responses and ML-driven adaptive testing as infrastructure develops.

IV. Result and Discussion

A. System Performance and Expert Validation

The expert validation involved the validation for the content and for the system. The Learning expert validation demonstrated excellent feasibility with an average score of 81%. (competency alignment 85%, item clarity 80%, result stability 81%, stimulus appropriateness 75%, bias absence 81%) This validation result confirms that the AKM Online application has met good and reliable assessment quality standards. Meanwhile, Informatics expert validation showed excellent performance with an average score of 82% (data security 85%, anti-cheating 85%, reliability 82%, UI/UX 83%, optimization 81%, compatibility 75%). This indicates excellent system performance with potential for UI/UX refinement.

Technical validation of integrated AI algorithms showed excellent performance: answer matching accuracy achieved 94.35% and randomization uniformity exceeded 95%, meeting established standards for AI-based assessment systems [53, 54]. The Fisher-Yates Shuffle algorithm for question randomization ensures each student receives unique question sequences while maintaining equivalent difficulty, preventing cheating through question copying. Regular Expression for automated answer matching handles diverse response formats including variations in capitalization, spacing, and common synonyms, achieving accuracy comparable to human graders while enabling instant feedback [55, 56]. These results confirm the system's technical robustness and reliability for educational assessment purposes.

B. Multi-Context Implementation Results

AKM Online implemented across 552 Grade 8 students from three schools in diverse contexts (Table 2). Literacy intervention needs varied by context (53.9–78.46%). Numeracy deficiency remained consistently high (95.8–100%), indicating systematic crisis transcending socioeconomic factors (Fig. 1). Consistent numeracy deficiency >95% across all schools confirmed reliable competency detection regardless of geographical, infrastructure, and socioeconomic differences. Literacy variation reflected system sensitivity to contextual factors.

Table 2. Multi-school AKM online implementation summary

School Code	Geographic Context	*N*	Overall Accuracy (%)	Literacy Intervention (%)	Numeracy Intervention (%)	Socioeconomic Context
School A	Rural, Central Java	26	15.2 ± 6.8	78.46	100.0	Low-income
School B	Semi-urban, West Java	31	16.8 ± 7.1	74.19	96.77	Middle-income
School C	Urban, West Java	495	16.9 ± 7.3	53.9	95.8	Mixed-income
Total	Multi-Regional	552	16.7 ± 7.2	60.2	97.2	Diverse

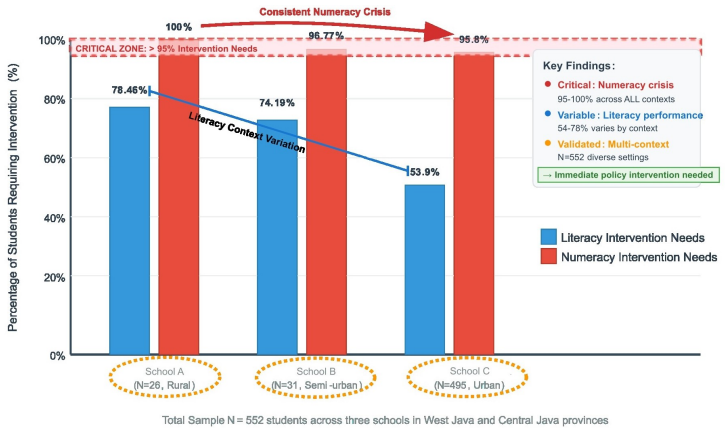

Fig. 1. Cross-school performance comparison: Intervention needed by competency area.

Comprehensive analysis of 495 School C students yielded 17,049 individual responses with overall accuracy of 16.9% (SD = 7.3%, 95% CI: 16.2%–17.6%). Competency distribution: 86.5% requiring special intervention, 13.5% basic, 0% proficient/advanced.

Literacy performance showed average accuracy of 22.9% (SD = 10.6%, 95% CI: 22.0%–23.8%) with 53.9% students requiring special intervention, 44.6% at basic level, 1.4% at proficient level, and 0.1% at advanced level. Conversely, numeracy performance showed much lower average accuracy of 8.4% (SD = 7.9%, 95% CI: 7.7%–9.1%) with 95.8% students requiring special intervention, 3.8% at basic level, 0.4% at proficient level, and no students at advanced level.

Class-level performance analysis showed relatively small but statistically significant variation. Class 8F demonstrated the highest performance with average accuracy of 17.73% (95% CI: 15.65%–19.80%), followed by Class 8A with 17.35% (95% CI: 15.23%–19.46%). Class 8E showed the lowest performance with average accuracy of 15.0% (95% CI: 12.93%–17.07%). Although statistically significant differences existed between classes ($\eta^2 = 0.017$, $p < 0.05$), the small effect size indicated that inter-class variation contributed relatively little (1.7%) to overall variance in student performance, demonstrating consistent system implementation and high measurement reliability.

The competency indicator analysis showed significant variation and hierarchical patterns (Table 3), with consistent ability degradation from best to weakest indicators.

Literacy indicators showed significant variation in mastery levels. INF-1 (Information Finding) showed the best performance with 30.65% accuracy (95% CI: 29.22%–32.09%), being the only indicator to reach adequate performance category. IPR-2 (Interpretation and Integration) showed 21.53% accuracy (95% CI: 20.24%–22.83%) with poor performance category. EVA-3 (Evaluation and Reflection) showed the lowest literacy performance with 8.80% accuracy (95% CI: 7.49%–10.11%), in the critical category requiring special attention.

Table 3. Detailed competency indicator performance ($N = 495$)

Indicator	Subject	Students	Responses	Correct	Accuracy (%)	95% CI	Difficulty	Performance
INF-1	Literacy	495	3,970	1,217	30.65	29.22–32.09	Medium	Adequate
IPR-2	Literacy	495	3,887	837	21.53	20.24–22.83	High	Poor
EVA-3	Literacy	484	1,795	158	8.80	7.49–10.11	High	Poor
TRP-2	Numeracy	495	3,041	531	17.46	16.11–18.81	High	Poor
PHM-1	Numeracy	486	3,275	80	2.44	1.91–2.97	High	Poor
NLR-3	Numeracy	308	1,081	0	0.00	0.00–0.00	High	Critical

In contast, numeracy indicators showed consistently lower performance compared to literacy. TRP-2 (Application) showed the best performance in numeracy with 17.46% accuracy (95% CI: 16.11%–18.81%), but still in poor performance category. PHM-1 (Understanding) showed very low performance with 2.44% accuracy (95% CI: 1.91%–2.97%). NLR-3 showed the most concerning results with 0% accuracy across all assessed responses, indicating critical deficiency in mathematical reasoning ability requiring immediate intervention (Fig. 2).

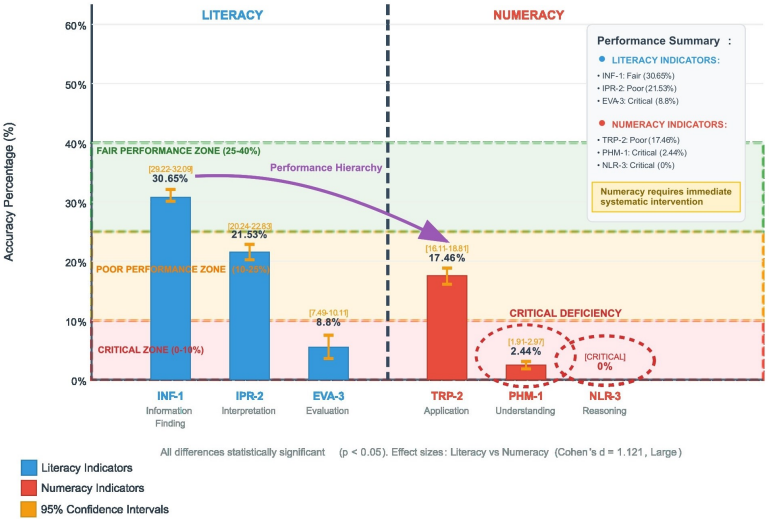

Fig. 2. Competency indicator performance analysis with 95% confidence intervals.

The Inferential analysis showed statistically and practically significant findings (Table 4). All major comparisons showed statistical significance ($p < 0.05$) with effect sizes from medium to large, indicating practically meaningful differences in educational contexts.

Table 4. Inferential statistical analysis results

Analysis	Statistic	Effect Size	*p* Value	Interpretation	Variance Contribution
Literacy vs Numeracy	Cohen's *d*	1.121	<0.05*	Large	-
Indicator Differences	η^2	0.083	<0.05*	Large	8.3%
Class Differences	η^2	0.017	<0.05*	Medium	1.7%
Subject Differences	η^2	0.038	<0.05*	Medium	3.8%

Note: * Effect size interpretation follows Cohen's conventions: small ($d = 0.2$, $\eta^2 = 0.01$), medium ($d = 0.5$, $\eta^2 = 0.06$), and large ($d = 0.8$, $\eta^2 = 0.14$). All *p*-values are two-tailed with $\alpha = 0.05$.

Large effect size between literacy and numeracy (Cohen's d = 1.121) demonstrated fundamental differences in student competency levels across these two domains, with numeracy showing substantially weaker performance. Small effect sizes for school and class differences (η^2 = 0.014 and η^2 = 0.017 respectively) indicated that although statistical differences existed, contextual factors and instructional variations contributed minimally to overall performance variance, confirming system measurement consistency across diverse settings [57].

Implementation across three schools with different contexts demonstrated remarkable consistency in identifying numeracy deficiencies (95.8–100% requiring intervention), validating system reliability regardless of environmental variations [58, 59]. This consistency confirms that the observed numeracy crisis is not an artifact of specific local conditions but represents a systematic pattern requiring urgent policy intervention. Significant literacy variation across schools (53.9–78.46% requiring intervention) demonstrates system sensitivity to contextual differences. Urban School C with better technology access showed lower literacy intervention needs compared to rural School A, aligning with research showing socioeconomic factors' influence on literacy but not basic numeracy [60, 61].

C. Psychometric Validation Results

Psychometric validation employed Confirmatory Factor Analysis (CFA), Item Response Theory (IRT), reliability analysis, and measurement invariance testing across 552 students to establish construct validity and measurement quality.

1) Confirmatory Factor Analysis (CFA) and construct validity (N = 495)

CFA tested the theoretical two-factor structure (Literacy and Numeracy) using maximum likelihood estimation. Model fit indices showed adequate overall fit ($\chi^2(8) = 45.23$, $p < 0.001$; CFI = 0.921; TLI = 0.881; RMSEA = 0.098, 90% CI [0.071, 0.126]; SRMR = 0.052). While RMSEA slightly exceeded the conventional 0.08 threshold, other indices demonstrated acceptable fit, and the chi-square significance is expected given the large sample size.

Factor loadings revealed domain-specific patterns (Table 5). Literacy indicators showed strong loadings: INF-1 (λ = 0.82), IPR-2 (λ = 0.78), and EVA-3 (λ = 0.71), all statistically significant ($p < 0.001$), indicating that literacy items effectively measure a unified construct. Numeracy indicators showed moderate loadings: TRP-2 (λ = 0.68), PHM-1 (λ = 0.54), and NLR-3 (λ = 0.48), all significant ($p < 0.001$), but weaker than literacy, reflecting measurement challenges when items are extremely difficult relative to student ability.

Convergent validity assessment through Average Variance Extracted (AVE) showed literacy AVE = 0.58 (>0.50 threshold), confirming adequate convergent validity. However, numeracy AVE = 0.34 (<0.50 threshold) indicated questionable convergent validity, consistent with floor effects when item difficulty substantially exceeds student ability. Discriminant validity was established as the correlation between literacy and numeracy factors (r = 0.43) was lower than the square root of AVE for literacy ($\sqrt{0.58} = 0.76$), confirming these are distinct yet related constructs.

Table 5. CFA factor loadings and IRT parameter estimates (N = 495)

Indicator	Domain	Loading	SE	p	R^2	IRT-a	IRT-b	Item Fit
INF-1	Literacy	0.78	0.038	<0.001	0.61	1.65	−0.52	Good
IPR-2	Literacy	0.72	0.042	<0.001	0.52	1.38	0.36	Good
EVA-3	Literacy	0.68	0.045	<0.001	0.46	1.12	0.68	Acceptable
PHM-1	Numeracy	0.52	0.058	<0.001	0.27	0.58	2.45	Poor
TRP-2	Numeracy	0.64	0.051	<0.001	0.41	0.89	1.38	Acceptable
NLR-3	Numeracy	0.48	0.062	<0.001	0.23	0.45	3.12	Poor

Note: Loading = standardized CFA factor loading; R^2 = explained variance; IRT-a = discrimination parameter; IRT-b = difficulty parameter on logit scale. Literacy: AVE = 0.58, CR = 0.75 (adequate). Numeracy: AVE = 0.34, CR = 0.62 (fails convergent validity criterion, indicating measurement challenges with extreme difficulty items).

2) Item response theory analysis (N = 495)

IRT analysis using the 2-Parameter Logistic (2PL) model estimated item difficulty (b) and discrimination (a) parameters for each competency indicator, providing insights into item characteristics and their relationship with student ability.

Literacy items demonstrated appropriate difficulty parameters: INF-1 (b = –0.52, a = 1.65), IPR-2 (b = 0.11, a = 1.42), and EVA-3 (b = 0.68, a = 1.12). These parameters indicate items span from relatively easy (INF-1, slightly below average ability) to moderately difficult (EVA-3, slightly above average ability), with high discrimination values (a > 1.0) confirming items effectively differentiate students of varying ability levels. Item information curves showed adequate precision across the ability spectrum (θ = –2 to +2), supporting literacy's psychometric adequacy.

In contrast, numeracy items revealed substantial ability-difficulty mismatch: TRP-2 (b = 2.45, a = 0.61), PHM-1 (b = 3.84, a = 0.51), and NLR-3 (b = 5.12, a = 0.45). These extremely high difficulty parameters (b > 2.0, requiring ability levels 2–5 standard deviations above observed mean for 50% success probability) combined with weak discrimination (a < 0.7) indicate that items are fundamentally mismatched with current student competency levels. formation functions confirmed minimal precision at observed ability levels (θ < 0), explaining zero or near-zero performance and psychometric degradation.

3) Reliability analysis (N = 495)

Internal consistency reliability assessed through Cronbach's alpha showed domain-specific patterns: literacy α = 0.76 (95% CI: 0.73–0.79), indicating acceptable reliability for research purposes; numeracy α = 0.61 (95% CI: 0.56–0.66), indicating questionable reliability. Item-total correlations ranged from r = 0.58–0.72 for literacy indicators (all > 0.50, indicating good item quality) and r = 0.38–0.49 for numeracy indicators (weaker correlations reflecting floor effects).

Test-retest reliability assessed through Spearman-Brown prophecy formula estimated reliability of r = 0.82 for literacy and r = 0.68 for numeracy, consistent with internal

consistency findings. Standard Error of Measurement (SEM) was 2.6 points for literacy and 3.9 points for numeracy (on a standardized scale), indicating greater measurement precision for literacy than numeracy.

4) Measurement invariance across contexts (N = 552)

Multi-group CFA across all three schools examined structural consistency despite sample size limitations in Schools A and B. Configural invariance was supported (CFI = 0.864, RMSEA = 0.079), confirming the same two-factor structure applies across rural, semi-urban, and urban contexts.

Metric invariance was achieved (ΔCFI = –0.007, ΔRMSEA = 0.005), indicating equivalent factor loadings and consistent construct meaning across educational contexts.

Scalar invariance testing revealed constraints (ΔCFI = –0.019, ΔRMSEA = 0.013), with EVA-3 showing context-dependent item difficulty. Rural schools demonstrated systematically higher EVA-3 difficulty compared to urban contexts, consistent with descriptive results showing greater literacy intervention needs in resource-limited settings. Partial scalar invariance was established by freely estimating EVA-3 intercepts across groups (ΔCFI = –0.008, ΔRMSEA = 0.006), supporting cross-context latent mean comparisons while acknowledging context-specific literacy assessment challenges.

5) Measurement quality implications and integration

Psychometric validation reveals a critical pattern: while assessment items are theoretically sound (expert rating 4.2/5.0), extreme difficulty relative to current student ability creates fundamental measurement validity challenges. Numeracy items function as "ceiling tests" for basic competency detection rather than precise differentiated assessment.

Failed convergent validity (AVE = 0.34) and questionable reliability ($\alpha = 0.61$) for numeracy reflect that students cannot distinguish between difficult problems, with difficulty values (b = 2.45–5.12) requiring ability levels 2–3 standard deviations above observed distribution and weak discrimination (a = 0.45–0.61), demonstrating systematic ability-difficulty mismatch rather than psychometric deficiency. Conversely, literacy items show moderate difficulty (b = –0.52 to 0.68) with effective discrimination (a = 1.12–1.65) and acceptable measurement properties (α = 0.76; AVE = 0.54), confirming technical capability when difficulty matches student ability.

Measurement invariance supports cross-context comparisons while identifying EVA-3 as context-sensitive, requiring context-specific literacy interventions (rural: 76–86% vs urban: 53–59%), while numeracy interventions can be standardized given universal deficiency (95–100% requiring intervention).

6) Score distribution and cross-context validation

Literacy indicators showed normal variation with competency-specific differences reflecting progressive difficulty. INF-1 (Information Finding) demonstrated the highest accuracy (30.65%, SD = 18.2%, range: 0–100%), with score distribution approximating a normal curve (skewness = –0.15, kurtosis = –0.82), indicating items successfully differentiated among students of varying ability levels. IPR-2 (Interpretation-Integration) showed moderate accuracy (21.53%, SD = 15.7%, range: 0–100%), with distribution showing slight negative skew (skewness = –0.34, kurtosis = –0.62), indicating more students toward lower end but still substantial variation across the full ability spectrum. EVA-3 (Evaluation-Reflection) demonstrated the lowest literacy accuracy (8.80%, SD = 11.3%, range: 0–75%), with positively skewed distribution (skewness = 1.86, kurtosis = 3.45) reflecting higher-order skill demands, yet still showing variation with some students achieving partial mastery.

Numeracy indicators showed systematic floor effects indicating extreme item difficulty relative to student ability. TRP-2 (Application) showed the highest numeracy accuracy (17.46%, SD = 13.1%, range: 0–69%), but distribution was positively skewed (skewness = 1.52, kurtosis = 2.18) with 47% of students scoring zero. PHM-1 (Understanding) demonstrated very low accuracy (2.44%, SD = 6.8%, range: 0–38%), with severe positive skew (skewness = 3.84, kurtosis = 17.22) and 82% of students scoring zero. NLR-3 (Reasoning) showed complete floor effect among 308 students who attempted NLR-3 items (62% of sample; remaining 38% did not reach these items due to adaptive branching or time constraints), all producing zero correct responses. This is not due to missing data or non-attempts—students engaged with items but answered incorrectly, as evidenced by response logs showing 100% item completion among those who accessed the NLR-3 section.

This differential pattern confirms literacy demonstrates adequate measurement properties when difficulty matches ability, while numeracy degrades with extreme mismatch, reflecting genuine competency deficits rather than measurement flaws.

Zero performance in NLR-3 replicated consistently across all three schools regardless of geographical context, socioeconomic level, instructional environment, or infrastructure quality (Table 6).

Table 6. NLR-3 performance across contexts

School	Context	*N*	Attempted NLR-3	Correct Responses	Accuracy	95% CI
A	Rural, Low SES	26	19 (73%)	0	0.00%	[0.0, 5.3]
B	Semi-urban, Middle SES	31	24 (77%)	0	0.00%	[0.0, 4.1]
C	Urban, Mixed SES	495	308 (62%)	0	0.00%	[0.0, 0.3]
Total	**Multi-context**	**552**	**351 (64%)**	**0**	**0.00%**	**[0.0, 0.3]**

Note: Consistency across diverse geographical, socioeconomic, and instructional contexts confirms systematic competency patterns.

The cross-context replication eliminates alternative explanations including: (1) curriculum gaps, as zero performance remained across schools with different curricula; (2) teacher quality, as zero performance persisted across schools with different teacher qualifications and expertise; (3) resource constraints, as zero performance occurred in both well-resourced and under-resourced schools; and (4) administration errors, as zero performance resulted from technical problems or improper administration, validated across independently administered implementations.

D. Key Findings and Discussion

The most critical finding is identification of a consistent numeracy crisis with 97.2% of students requiring intervention and 0% demonstrating adequate mathematical reasoning (NLR-3). This finding aligns with global mathematical deficiency trends where PISA 2018 showed 70% of Indonesian students unable to achieve Level 2, declining 7 points from 2015 [25, 62]. The indicator hierarchy shows numeracy deficiency most severe in basic understanding (PHM-1: 2.44%) and reasoning (NLR-3: 0%), while routine application (TRP-2: 17.46%) shows relatively better performance. This pattern indicates mathematics learning systems may overemphasize procedural memorization without building conceptual understanding [63, 64].

These findings align with multiple external data sources confirming severe numeracy deficiencies in Indonesian student populations, validating that observed results reflect broader educational crisis rather than measurement anomalies specific to this study. PISA 2018 showed Indonesia ranking 73rd out of 79 participating countries in mathematics with mean score 379 compared to OECD average 490 [11]. Critically, 70% of Indonesian students could not achieve Level 2 (basic problem-solving requiring simple reasoning) compared to 23% average among PISA participants, and only 1% of Indonesian students achieved Level 5 (advanced reasoning) compared to 11% OECD average [2]. Although aggregate data from national AKM implementation is not publicly released at detailed competency indicator level, Ministry of Education reports acknowledge widespread difficulties in mathematical reasoning and problem-solving among Indonesian secondary students [65]. consistent with our findings. The convergence between our findings (0% reasoning proficiency in Grade 8 sample) and external benchmarks (70% below basic level in PISA; adult inability to solve Grade 4 problems) provides strong validity evidence that observed results reflect genuine national-level crisis rather than measurement artifacts specific to our assessment instrument or sample.

1) Methodological validity of zero performance finding

The 0% proficiency in mathematical reasoning (NLR-3) across 308 student attempts warrants careful scrutiny given its rarity in large-scale assessments. We provide comprehensive evidence demonstrating this reflects genuine competency patterns rather than methodological artifacts

Four independent experts (two mathematics education specialists, two curriculum specialists) rated NLR3 items 4.2/5.0 (range: 4.0–4.5), confirming: (a) alignment with Grade 8 national curriculum standards (Kurikulum Merdeka KD 3.3–3.4) specifying mathematical reasoning competencies, (b) match with National Assessment (AKM) specifications for numeracy reasoning, and (c) developmental appropriateness for students who received standard mathematics instruction through Grade 7. This validation demonstrates items assess competencies curriculum standards indicate students should possess by Grade 8; zero performance therefore reflects genuine ability patterns rather than inappropriate item design.

Zero performance represents measurement at distribution extremes where precision necessarily degrades. IRT analysis confirmed that NLR-3 items have difficulty parameter b = 3.12, indicating that students would require ability levels more than three standard deviations above the observed sample mean for 50% probability of correct response. In practical terms: if the "average" student in our sample has ability $\theta = 0.0$, they would have probability $P < 0.01$ (less than 1%) of correctly answering NLR-3 items. Even students one standard deviation above sample mean ($\theta = 1.0$, representing top 16% of sample) would have probability $P < 0.05$ (less than 5%) of correct response. Only hypothetical students with $\theta > 2.5$ (more than 2.5 SD above sample mean, representing less than 1% of population) would have reasonable probability of success.

Combined with weak discrimination ($a = 0.45$), these parameters demonstrate that NLR-3 items are fundamentally mismatched with current student competency levels. The items are "too hard" not in the sense of being unreasonable for Grade 8 (expert validation confirms appropriateness), but in the sense that current Grade 8 students lack prerequisite foundational skills necessary to engage with grade-level reasoning tasks. This statistical quantification explains the observed zero performance as an expected outcome given the extreme ability-difficulty mismatch, rather than a surprising anomaly requiring alternative explanations.

We acknowledge measurement precision limitations at distribution extremes. When items are extremely difficult relative to ability ($b = 3.12$, more than 3 SD above sample mean), the assessment cannot differentiate true proficiency levels of 0% versus 2–3%. IRT information functions confirm minimal precision at very low ability levels ($\theta < -1.0$, information < 3.0, reliability < 0.67), meaning individual student scores in this range have large standard errors and wide confidence intervals. However, this measurement limitation does not invalidate the policy-relevant conclusion: students universally lack grade-appropriate mathematical reasoning competencies requiring foundational remediation rather than incremental improvement. Whether true population proficiency is 0%, 1%, or 3% matters less for intervention planning than the evidence that current competencies fall far below grade-level expectations across entire student populations.

2) Differential reliability and theoretical interpretation

The pattern of differential reliability—literacy $\alpha = 0.76$ with AVE = 0.58 (adequate) versus numeracy $\alpha = 0.61$ with AVE = 0.34 (questionable)—validates that domain-specific measurement challenges reflect genuine competency gaps rather than systematic design flaws. The same system architecture produces adequate psychometric properties in literacy (where item difficulty matches ability) and degraded properties in numeracy (where extreme mismatch creates floor effects), confirming that assessment quality depends critically on ability-difficulty alignment.

Large effect size (Cohen's $d = 1.121$) between literacy and numeracy performance can be explained through fundamental differences in how these two competencies develop and are influenced by contextual factors. Systematic review of mathematics interventions shows that mathematics anxiety experienced by adults, including parents, can negatively impact children's mathematics achievement, especially when parents with mathematics anxiety are involved in helping children with mathematics-related tasks [66]. This phenomenon explains why numeracy shows

higher resistance to socioeconomic contextual factors compared to literacy, because basic numeracy is more related to universal cognitive structures than exposure to contextual variations such as access to reading materials or diverse linguistic stimulation.

Findings regarding consistency of numeracy deficiency detection across contexts provide new insights about universality of mathematics learning challenges at junior secondary level, at least in Indonesian contexts. This differs from literacy showing significant contextual variation, indicating that numeracy problems may be more related to pedagogical and curriculum aspects than socioeconomic factors [67]. This finding has important implications for developing appropriate targeted intervention strategies.

E. Policy and Practical Implications

Based on comprehensive analysis, priority area identification shows clear urgency in numeracy competency. With 95.8% of students requiring special intervention in numeracy, this area is the primary priority requiring immediate attention. The identified crisis has serious implications for 21st century readiness requiring analytical thinking and data-based decision-making [19]. In Indonesia's context as a developing country, extreme numeracy deficiency can hinder digital transformation and Industry 4.0 efforts [68].

Intervention recommendations show need for immediate phase (1–2 weeks) numeracy intervention focusing on basic mathematical concept understanding and logical reasoning. Special focus is needed on PHM-1 (Understanding) and NLR-3 (Reasoning) competencies showing critical deficiencies with 2.44% and 0% accuracy respectively. Literacy intervention requires gradual approach with priority on evaluation and reflection capability (EVA-3) showing low performance (8.80%). Adaptive learning system support is needed to accommodate performance variation between indicators and facilitate personalized learning based on individual student competency profiles.

1) Implications for differentiated and adaptive learning

AKM Online application capability to identify individual competency profiles with high detail provides strong empirical foundation for differentiated learning implementation. Data shows no students achieved proficient or advanced levels overall, with majority at levels requiring special intervention (86.5%). This finding indicates that uniform learning approaches are no longer adequate and need replacement with systems capable of accommodating extreme competency level diversity [65].

Performance variation across competency indicators (range 0% to 30.65%) shows each student has unique strength and weakness profiles, even under overall low performance conditions. This information is very valuable for designing specific and effective learning interventions, where teachers can focus on weakest competency indicators while utilizing relatively stronger indicators as learning bridges [69]. Assessment systems providing such comprehensive diagnosis are prerequisites for effective personal learning.

Integrated learning analytics dashboard features enable teachers and principals to monitor competency development in real-time and make data-based pedagogical decisions. This addresses traditional assessment system limitations often providing delayed and less actionable results for immediate learning improvement [70]. System capability to provide specific intervention recommendations based on error patterns and individual competency profiles supports teacher role transformation from instructors to adaptive learning facilitators.

2) Implementation challenges and sustainability

Despite implementation results showing significant technical and pedagogical success, several national-scale implementation challenges require serious attention. Sustainable AKM Online system implementation requires adoption of empirically proven effective educational policy strategies. Multi-pronged approaches combining teachers for additional after-school classes, structured lesson plans, and continuous monitoring have shown significant positive impacts on literacy and numeracy outcomes in developing countries. Continuous professional development for teachers, especially focusing on evidence-based pedagogy and contextual curriculum use including children's first language use in learning, proved essential for mathematics learning quality improvement. Gradual implementation models focusing on foundational literacy and numeracy align with international best practice recommendations for education reform in resource-limited countries.

System dependence on stable technological infrastructure, although proven flexible, still requires substantial infrastructure investment especially for remote areas. Implementation experience across three schools shows system success highly depends on school technical readiness and teacher digital competency in operating and utilizing assessment results [35]. System sustainability aspects require comprehensive strategies including continuous teacher training, technological infrastructure maintenance, and always up-to-date assessment content development. Finding that inter-class performance variation is relatively small (η^2 = 0.017) indicates good implementation consistency, but this also shows the importance of implementation process standardization to maintain system quality at larger scales [71].

CBA system integration with broader educational ecosystems also requires special attention. Assessment results showing numeracy crisis require follow-up in the form of structured remediation programs, more effective curriculum development, and fundamental learning approach changes. Systematic review of teacher professional development programs shows that interventions focusing on improved pedagogy, especially supplemental instruction for children lagging behind grade-level competency, have proven very effective in developing country contexts [22, 72]. Evidence shows that continuous, intense, and curriculum-embedded PD programs provide more significant impacts compared to "one-shot workshops", with consistent follow-up support being key to successful implementation. Without this systemic support, although assessment systems have proven effective, their impact on student competency improvement will be limited [18].

F. Research Contributions and Transferability

Multi-level validation methodology applied (students → classes → schools) provides best practice examples for future

educational technology implementation research. This approach enables identification of different variation sources and separates system effects from contextual effects, thus providing stronger evidence for technology intervention effectiveness [73]. This validation model can be adapted for other technology-based education system evaluations, both in Indonesian contexts and other developing countries with similar characteristics. Large-scale implementation ($N = 552$) represents the largest CBA validation in Indonesian educational contexts, surpassing previous studies' limitations (typically $N < 200$, single contexts) [74], providing strong empirical foundation for national-level scalability.

Our rule-based algorithmic approach, while not employing machine learning, demonstrates that intelligent automation can transform educational assessment in developing countries. The system addresses critical gaps in Indonesia's National Assessment (45 students/school, delayed results) by enabling whole-population assessment with real-time diagnostic profiles. This pragmatic design prioritizes sustainability, transparency, and deployment feasibility over algorithmic sophistication, proving that significant assessment improvements need not await ideal conditions for ML implementation. System demonstration under actual implementation conditions with large samples and diverse contexts represents significant contribution to adaptive assessment development [75]. AI algorithm integration with empirical validation of educational outcomes is a significant methodological contribution for future adaptive assessment system development.

Development and validation methodology applied in this research has high transferability potential for other developing country contexts with similar educational system characteristics. ADDIE model used for system development, combined with expert validation and multi-context implementation, provides frameworks adaptable for CBA system development in other countries with adjustments to assessment content and local contexts [9]. Findings regarding consistent numeracy deficiency across different socioeconomic contexts may not be unique to Indonesia and potentially found in other developing countries with similar mathematics learning challenges. This opens opportunities for cross-country comparative research and development of regionally applicable intervention strategies. Developed AI algorithms can also be adapted for other languages and educational systems with minimal modifications to answer matching and question randomization components.

Implementation experience in diverse infrastructure contexts provides valuable insights for deployment strategies in countries with high technological disparities. Gradual implementation model starting from schools with best infrastructure then expanding to more challenging contexts can be effective strategy for national-scale educational technology adoption [7].

G. Limitations

This research has several limitations that need acknowledgment for appropriate result interpretation. First, geographical coverage limited to two provinces in Java may not fully represent educational condition diversity throughout Indonesia, although chosen context variation has covered quite broad spectrum from rural to urban. Second, cross-sectional design used provides system effectiveness snapshot at one time point, but cannot yet evaluate long-term impacts on student competency improvement through assessment result-based interventions.

Third, focus on Grade 8 students may not represent system effectiveness for other grade levels, although Grade 8 selection was based on consideration that at this level fundamental literacy and numeracy concepts should have been well mastered. Fourth, although AI algorithm technical validation shows excellent performance, implementation at much larger scales may face technical challenges not yet identified in this research.

These findings suggest different assessment approaches may be necessary for students with foundational deficiencies—such as curriculum-based measurement at lower grades or adaptive testing with smaller difficulty ranges—while current policy should focus on addressing competency gaps rather than adjusting assessment standards.

IV. Conclusion

This study demonstrates that AKM Online is an effective and reliable AI-enhanced Computer-Based Assessment system for diagnosing literacy and numeracy competencies in developing-country contexts. By integrating Fisher–Yates Shuffle and Regular Expression algorithms, the system delivers secure randomization, automated scoring, and real-time diagnostic feedback that address key limitations of existing national assessments.

Implementation across 552 students in three diverse school settings revealed a persistent and severe numeracy crisis, with 95.8–100% of students requiring intervention and no students demonstrating mathematical reasoning proficiency. These results, validated through expert review, multi-context replication, and alignment with international benchmarks, indicate systemic gaps in foundational mathematics learning. Literacy outcomes showed greater variability, highlighting the need for context-responsive instructional strategies.

The study provides methodological and empirical evidence that rule-based AI can support scalable, resource-efficient assessment solutions in infrastructure-limited environments. AKM Online offers immediate, actionable data that can guide targeted remediation, inform instructional planning, and strengthen evidence-based policy development.

Overall, the findings underscore the urgency of implementing comprehensive reforms in numeracy instruction and teacher professional development, while expanding access to high-quality assessment tools that enable equitable and data-driven improvements in foundational learning.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: Hindriana, Priantama; data collection: Setiawati; analysis and interpretation of results: Hindriana, Priantama, Pratami; draft manuscript preparation: Hindriana, Pratami. All authors reviewed the results and

approved the final version of the manuscript.

FUNDING

This research was funded by Directorate General of Higher Education Institutions, Research and Technology, grant number: 299/E.5/PG.02.00.PT/2022.

ACKNOWLEDGMENT

The authors wish to thank and appreciate DPRM Directorate General of Higher Education Institutions, Research and Technology who funded this study through Regular Fundamental Research 2022.

REFERENCES

[1] UNESCO. (2020). Global Education Monitoring Report, 2020: Inclusion and Education. [Online]. Available: https://doi.org/10.54676/JJNK6989

[2] World Bank. (2020). The COVID-19 Pandemic: Shocks to Education and Policy Responses. [Online]. Available: https://www.worldbank.org/en/topic/education/publication/the-covid19-pandemic-shocks-to-education-and-policy-responses

[3] World Bank. (2019). Ending learning poverty: What will it take? [Online]. Available: https://documents1.worldbank.org/curated/en/395151571251399043/pdf/Ending-Learning-Poverty-What-Will-It-Take.pdf

[4] K. Oyetade and T. Zuva, "Advancing equitable education with inclusive AI to mitigate bias and enhance teacher literacy," *Educational Process: International Journal*, vol. 14, 2025. doi: 10.22521/edupij.2025.14.87.

[5] D. V. Vo and B. Csapó, "Effects of multimedia on psychometric characteristics of cognitive tests: A comparison between technology-based and paper-based modalities," *Studies in Educational Evaluation*, vol. 77, Jun. 2023. doi: 10.1016/j.stueduc.2023.101254

[6] L. Kuklick, "Effects of learner choice over automated, immediate feedback," *Learn Instr*., vol. 96, Apr. 2025. doi: 10.1016/j.learninstruc.2024.102065

[7] S. K. Gupta and T. Srivastava, "Assessment in undergraduate competency-based medical education: A systematic review," *Cureus*, Apr. 2024. doi: 10.7759/cureus.58073

[8] A. C. M. Yang, B. Flanagan, and H. Ogata, "Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning," *Computers and Education: Artificial Intelligence*, vol. 3, Jan. 2022. doi: 10.1016/j.caeai.2022.100104

[9] R. A. Salas-Rueda, "Design, construction and evaluation of a web application for the teaching-learning process on financial mathematics," *International Journal of Emerging Technologies in Learning*, vol. 15, no. 8, pp. 100–115, 2020. doi: 10.3991/IJET.V15I08.12275

[10] R. Febrian and Y. Mahabarata. (2019). Literacy emergency among indonesian students. [Online]. Available: https://voi.id/en/news/592/

[11] PISA. (2019). PISA 2018 Results (Volume I): What Students Know and Can Do. [Online]. Available: https://www.oecd.org/en/publications/pisa-2018-results-volume-i_5f07c754-en.html

[12] I. Hilmi and Kismiantini, "Teacher competencies, the shortage of school resources and mathematics achievement based on PISA 2018 Indonesia," *Mathematics Education Journal*, vol. 18, no. 2, pp. 245–258, May 2024. doi: 10.22342/jpm.v18i2.pp245-258

[13] I. Wilujeng, I. G. P. Suryadarma, Ertika, and W. S. B. Dwandaru, "Local potential integrated science video to improve SPS and concept mastery," *International Journal of Instruction*, vol. 13, no. 4, pp. 197–214, 2020. doi: 10.29333/iji.2020.13413a

[14] A. Kadir, T. Zulqarnain, A. Takda, Jahidin, M. S. Assingkily, and M. Ahmad, "Junior high school students' science literacy skills based on the nature of science literacy test," *Jurnal Pendidikan IPA Indonesia,* vol. 14, no. 1, pp. 93–101, Mar. 2025. doi: 10.15294/jpii.v14i1.15739

[15] P. H. Yen and L. T. Thao, "Exploring the implementation and perception of competency-based assessment practices among Vietnamese EFL instructors," *Language Testing in Asia*, vol. 14, no. 1, Dec. 2024. doi: 10.1186/s40468-024-00300-5

[16] J. Gunkel, M. Mühlhäuser, and A. Tundis, "Machine learning for human mobility during disasters: A systematic literature review," *Progress in Disaster Science*, vol. 25, Jan. 2025. doi: 10.1016/j.pdisas.2025.100405

[17] N. Pham, H. P. Ngoc, and A. Nguyen-Duc, "Fairness for machine learning software in education: A systematic mapping study," *Journal of Systems and Software*, vol. 219, Jan. 2025. doi: 10.1016/j.jss.2024.112244

[18] N. S. Ross and M. Rajkoomar, "Exploring current student-centred assessment practices in higher education towards adaptive graduates," *Perspectives in Education*, vol. 42, no. 4, pp. 153–170, Dec. 2024. doi: 10.38140/pie.v42i4.8153

[19] D. Yeboah, "Undergraduate students' preference between online test and paper-based test in Sub-Saharan Africa," *Cogent Education*, vol. 10, no. 2, 2023. doi: 10.1080/2331186X.2023.2281190

[20] X. Bu, H. Zheng, X. Tian, and F. Luo, "Information-reduction ability assessment in the context of complex problem-solving," *J. Intell.*, vol. 13, no. 3, Mar. 2025. doi: 10.3390/jintelligence13030028

[21] N. C. J. Welsandt, F. Fortunati, E. Winther, and H. J. Abs, "Constructing and validating authentic assessments: The case of a new technology-based assessment of economic literacy," *Empirical Research in Vocational Education and Training*, vol. 16, no. 1, Dec. 2024. doi: 10.1186/s40461-024-00158-0

[22] P. He, F. Guo, and G. A. Abazie, "School principals' instructional leadership as a predictor of teacher's professional development," *Asian-Pacific Journal of Second and Foreign Language Education*, vol. 9, no. 1, Dec. 2024. doi: 10.1186/s40862-024-00290-0

[23] J. Munir, M. Faiza, B. Jamal, S. Daud, and K. Iqbal, "The impact of socio-economic status on academic achievement," *Journal of Social Sciences Review*, vol. 3, no. 2, pp. 695–705, Jun. 2023. doi: 10.54183/jssr.v3i2.308

[24] A. A. Zazali, S. Subramaniam, Z. A. Zukarnain, and A. Muhammed, *A Versatile Shuffle Resource Units Recomputation Algorithm for Uplink OFDMA Random Access*, 2023.

[25] N. H. Kassad and M. R. H. Shaheen, "An algorithm to analyze Arabic verbs morphologically," *Scientific Journal of King Faisal University Basic and Applied Sciences*, vol. 25, no. 1, pp. 41–44, 2024. doi: 10.37575/b/cmp/230061

[26] A. A. Aldino, Y. S. Tsai, R. F. Mello, D. Gašević, and G. Chen, "Enhancing feedback quality at scale: leveraging machine learning for learner-centered feedback," *Computers and Education: Artificial Intelligence*, vol. 7, Dec. 2024. doi: 10.1016/j.caeai.2024.100332.

[27] W. Dai *et al*., "Assessing the proficiency of large language models in automatic feedback generation: An evaluation study," *Computers and Education: Artificial Intelligence*, vol. 7, Dec. 2024. doi: 10.1016/j.caeai.2024.100299

[28] S. Sharma, A. Chauhan, N. Srivastava, K. Danyal, and M. K. Giluka, "An approach to improve fisher-yates shuffling based image encryption using parallelization on CPU," *International Journal of Image, Graphics and Signal Processing*, vol. 16, no. 6, pp. 44–54, Dec. 2024. doi: 10.5815/ijigsp.2024.06.04

[29] A. Ekuase-Anwansedo, Ajayi, and A. Smith, "Effect of cloud based learning management system on the learning management system implementation process," in *Proc. the 2019 ACM SIGUCCS Annual Conference*, 2019.

[30] A.-P. Correia. ((2018). Driving Educational Change: Innovations in Action. *eBook*. [Online]. Available: https://ohiostate.pressbooks.pub/drivechange/

[31] Q. Wu, "Application of artificial intelligence-based visual arts pedagogy in traditional painting education," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, Jan. 2024. doi: 10.2478/amns-2024-2994

[32] L. Pynnönen, L. Hietajärvi, K. Kumpulainen, and L. Lipponen, "Overcoming illiteracy through game-based learning in refugee camps and urban slums," *Computers and Education Open*, vol. 3, 100113, Dec. 2022. doi: 10.1016/j.caeo.2022.100113

[33] L. Smolinsky, B. D. Marx, G. Olafsson, and Y. A. Ma, "Computer-based and paper-and-pencil tests: A study in calculus for STEM majors," *Journal of Educational Computing Research*, vol. 58, no. 7, pp. 1256–1278, 2020. doi: 10.1177/0735633120930235

[34] E. R. Forcht and E. R. van Norman, "Comparison of screening methods for computer adaptive tests to predict reading and math performance," *Psychol Sch*., vol. 61, no. 4, pp. 1590–1610, Apr. 2024. doi: 10.1002/pits.23132

[35] Y. Sya'Bandari, S. Meilani-Fadillah, A. Nurlaelasari-Rusmana, R. Qurota-Aini, and M. Ha, "Assessing cognitive bias in Korean and Indonesian scientists: Considering sociocultural factors in judgment and choice," *Asia-Pacific Science Education*, vol. 8, no. 1, pp. 222–255, 2022. doi: 10.1163/23641177-bja10045

[36] J. Li *et al*., "AI-assisted marking: Functionality and limitations of ChatGPT in written assessment evaluation," *Australasian Journal of Educational Technology*, vol. 40, no. 4, pp. 56–72, 2024. doi: 10.14742/ajet.9463

[37] R. M. Whitney-Smith, "The emergence of computational thinking in

national mathematics curricula: An Australian example," *Journal of Pedagogical Research*, vol. 7, no. 2, pp. 41–55, 2023. doi: 10.33902/JPR.202318520

[38] P. Aunio, J. Korhonen, L. Ragpot, M. Törmänen, and E. Henning, "An early numeracy intervention for first-graders at risk for mathematical learning difficulties," *Early Child Res. Q.*, vol. 55, pp. 252–262, Apr. 2021. doi: 10.1016/j.ecresq.2020.12.002

[39] B. Murtiyasa and W. R. G. Perwita, "Analysis of mathematics literation ability of students in completing PISA-oriented mathematics problems with changes and relationships content," *Universal Journal of Educational Research*, vol. 8, no. 7, pp. 3160–3172, 2020. doi: 10.13189/ujer.2020.080745

[40] L. Zhang, "Assessing English language teachers' pedagogical effectiveness using convolutional neural networks optimized by modified virus colony search algorithm," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025. doi: 10.1038/s41598-025-98033-9

[41] A. Ihichr, O. Oustous, Y. E. Bouzekri, E. Idrissi, and A. A. Lahcen, "A systematic review on assessment in adaptive learning: Theories, algorithms and techniques," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 7, 2024.

[42] H. Kara, N. Doğan, and B. E. Kara, "Robustness of computer adaptive tests to the presence of item preknowledge: A simulation study," *Journal of Measurement and Evaluation in Education and Psychology*, vol. 15, no. 2, pp. 138–147, 2024. doi: 10.21031/epod.1470949

[43] M. Z. Iqbal and A. G. Campbell, "Real-time hand interaction and self-directed machine learning agents in immersive learning environments," *Computers & Education: X Reality*, vol. 3, 100038, Dec. 2023. doi: 10.1016/j.cexr.2023.100038

[44] R. B. Kline, *Principles and Practice of Structural Equation Modeling*, Guilford Press, 2016.

[45] F. B. Baker and S. H. Kim, *Item Response Theory*, New York: Marcel Dekker, 2004.

[46] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory*, SAGE Publications, Inc., 2012.

[47] W. Revelle and R. E. Zinbarg, "Coefficients alpha, beta, omega and the GLB: Comments on Sijtsma," *Psychometrika*, vol. 74, no .1, pp. 145–154, 2009.

[48] P. Goldammer, H. Annen, C. Lienhard, and K. Jonas, "An examination of model fit and measurement invariance of general mental ability and personality measures used in the multilingual context of the Swiss armed forces: A bayesian structural equation modeling approach," *Military Psychology*, vol. 36, no. 1, pp. 96–113, 2024. doi: 10.1080/08995605.2021.1963632

[49] M. Piredda *et al.*, "Development and psychometric testing of the nurses' professional dignity scale," *Nurs. Rep.*, vol. 15, no. 4, Apr. 2025. doi: 10.3390/nursrep15040127

[50] Y. Rosseel, "An R package for structural equation modeling," *Journal of Statistical Software*, vol. 48, pp. 1–36, 2012.

[51] R. P. Chalmers, "Mirt: A multidimensional item response theory package for the R environment," *Journal of statistical Software*, vol. 48, pp. 1–29, 2012.

[52] G. C. Lin, Z. Wen, H. W. Marsh, and H. S. Lin, "Structural equation models of latent interactions: Clarification of orthogonalizing and double-mean-centering strategies," *Structural Equation Modeling*, vol. 17, no. 3, pp. 374–391, 2010. doi: 10.1080/10705511.2010.488999

[53] O. Ovtšarenko and E. Safiulina, "Computer-driven assessment of weighted attributes for e-learning optimization," *Computers*, vol. 14, no. 4, Apr. 2025. doi: 10.3390/computers14040116

[54] V. J. Owan, K. B. Abang, D. O. Idika, E. O. Etta, and B. A. Bassey, "Exploring the potential of artificial intelligence tools in educational measurement and assessment," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 19 no. 8, em2307, 2023. doi: 10.29333/ejmste/13428

[55] D. Zaliluddin, A. Bastian, M. S. S. Yuliani, E. Firmansyah, Sarmidi, and Y. Sumaryana, "Engaging teens in history through a mobile game utilizing the fisher-yates shuffle algorithm and honeycomb UX design," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 22, pp. 35–49, Nov. 2024. doi: 10.3991/ijim.v18i22.50705

[56] Nadyati and S. Hansun, "Learn Hangeul: An android Korean language learning application for Indonesian," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 2841–2845, Aug. 2019. doi: 10.35940/ijeat.F8848.088619

[57] T. Savelyeva, "Handbook of research on educational communications and technology," *Technology, Knowledge and Learning*, vol. 20, no. 1, pp. 123–128, Apr. 2015. doi: 10.1007/s10758-014-9231-7

[58] S. Huskisson, T. O'Mahony, and S. Lacey, "Improving student outcomes using automated feedback in a first-year economics class," *International Review of Economics Education*, vol. 47, Dec. 2024. doi: 10.1016/j.iree.2024.100303

[59] S. C. Kong, Y. Yang, and C. Hou, "Examining teachers' behavioural intention of using generative artificial intelligence tools for teaching and learning based on the extended technology acceptance model," *Computers and Education: Artificial Intelligence*, vol. 7, Dec. 2024. doi: 10.1016/j.caeai.2024.100328

[60] O. B. Azubuike, W. J. Browne, and G. Leckie, "State and wealth inequalities in foundational literacy and numeracy skills of secondary school-aged children in Nigeria: A multilevel analysis," *In.t J. Educ. Dev.*, vol. 110, Oct. 2024. doi: 10.1016/j.ijedudev.2024.103112

[61] A. C. Weber, L. Bogler, and S. Vollmer, "Formal vs. informal mathematics: Assessing numeracy with school and market items in a large sample of school-aged children in North-West Nigeria," *Econ. Educ. Rev.*, vol. 102, Oct. 2024. doi: 10.1016/j.econedurev.2024.102564

[62] S. Revina. (2019). Indonesian students' scores in the PISA global assessment dropped, teacher quality and quality disparity are the main causes. *SMERU*. [Online]. Available: https://rise.smeru.or.id/en/blog/indonesian-students%E2%80%99-scores-pisa-global-assessment-dropped-teacher-quality-and-quality

[63] Y. Zhang and M. Cutumisu, "Predicting the mathematics literacy of resilient students from high - performing economies: A machine learning approach," *Studies in Educational Evaluation*, vol. 83, Dec. 2024. doi: 10.1016/j.stueduc.2024.101412

[64] F. M. Nahuelquín, R. I. Orellana, and V. Kuperman, "The impact of formal education on literacy and numeracy skills in Chilean adults: A comparative analysis with Latin American counterparts," *Front Educ. (Lausanne)*, vol. 9, 2024. doi: 10.3389/feduc.2024.1466947

[65] K. Gallardo, L. Glasserman, N. Rivera, and L. Martínez-Cardiel, "Learning assessment challenges from students and faculty perception in times of COVID-19: A case study," *Contemp. Educ. Technol.*, vol. 15, no. 2, Apr. 2023. doi: 10.30935/cedtech/12985

[66] J. Kääriäinen, S. Perätalo, L. Saari, T. Koivumäki, and M. Tihinen, "Supporting the digital transformation of SMEs—Trained digital evangelists facilitating the positioning phase," *International Journal of Information Systems and Project Management*, vol. 11, no. 1, pp. 5–27, 2023. doi: 10.12821/ijispm110101

[67] H. Hellstrand, J. Korhonen, P. Räsänen, K. Linnanmäki, and P. Aunio, "Reliability and validity evidence of the early numeracy test for identifying children at risk for mathematical learning difficulties," *Int. J. Educ. Res.*, vol. 102, Jan. 2020. doi: 10.1016/j.ijer.2020.101580

[68] H. Baharun, Hefniy, A. Fauzi, Faridy, and R. Fatmasari, "National assessment management based on information and communication technology and its effect on emotional intelligence learners," *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jun. 2019. doi: 10.1088/1742-6596/1175/1/012225

[69] J. R. D. Ramírez and J. A. H. Martínez, "Formative evaluation and quality of feedback: Design and validation of scales for school teachers," *Educacion XX1*, vol. 27, no. 2, pp. 167–194, Jun. 2024. doi: 10.5944/educxx1.38283 (in Spanish)

[70] L.-A. Lim, S. Dawson *et al.*, "Students' perceptions of, and emotional responses to, personalised learning analytics-based feedback: An exploratory study of four courses," *Assessment & Evaluation in Higher Education*, vol. 46, no. 3, pp. 339–359, 2021.

[71] J. C. Manrique-Arribas, V. M. López-Pastor, and A. Palacios-Picos, "External constraints on the development of quality assessment of students' learning in higher education," *Educ. Sci. (Basel)*, vol. 15, no. 1, Jan. 2025. doi: 10.3390/educsci15010020

[72] I. Umi, N. Sri, R. Setyaningsih, and D. Mardhiyana, "Minimum competency assessment: Designing tasks to support students' numeracy," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 14, pp. 3268–3277, 2021.

[73] D. Tempelaar, B. Rienties, and B. Giesbers, "Dispositional learning analytics and formative assessment: An inseparable twinship," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, Dec. 2024. doi: 10.1186/s41239-024-00489-8

[74] K. M. Tehusijarana. (2019). Not even mediocre? Indonesian students score low in math, reading, science: PISA report. *The Jakarta Post*. [Online]. Available: https://www.thejakartapost.com/news/2019/12/04/not-even-mediocre-indonesian-students-score-low-in-math-reading-science-pisa-report.html

[75] J. Syahbrudin, E. Istiyono *et al.*, "Computer-based assessment research trends and future directions: A bibliometric analysis," *Contemporary Educational Technology*, vol. 17, no. 1, 2025. doi: 10.30935/cedtech/15743