

A Systematic Review of the ARGRA Model for Technology-Based Reading Diagnosis in Elementary Education

Ita Kurnia^{1*}, Alfi Laila², Saktian Dwi Hartantri³

itakurnia@unpkediri.ac.id¹, alfilaila@unpkediri.ac.id², saktiandwihartantri@gmail.com³

Elementary School Teacher Education, FKIP, Universitas Nusantara PGRI Kediri^{1,2}

Elementary School Teacher Education, FKIP, Universitas Muhammadiyah Tangerang³

ARTICLE INFO

Article history:

Received June 20th 2026

Accepted July 24th, 2026

Available online July 30th 2026

Kata kunci:

Asesmen Membaca Adaptif, Analisis Keterbacaan, Gamifikasi, Android, Diagnosis Level Membaca, Sekolah Dasar

Keywords:

Adaptive Reading Assessment, Readability Analysis, Gamification, Android, Reading Level Diagnosis, Elementary School

This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).

Copyright © 2026 by Author

Published by:

[Universitas Nusantara PGRI Kediri](https://unpkediri.ac.id)

Email: jurnal_jpdn@unpkediri.ac.id



Abstract :

Accurate reading assessment is fundamental to elementary school literacy development. The ARGRA model (Adaptive Reading Assessment, Readability Analysis, Gamification, and Android-Based Assessment) addressed the static nature and poor engagement of conventional testing. Following PRISMA 2020 guidelines,

a systematic review was conducted across six databases (2009—2025). Out of 38 full-text articles assessed, 30 independent empirical studies (reported across 37 papers) met inclusion criteria. Methodological quality was evaluated using the Newcastle Ottawa Scale (NOS), categorizing 78% of studies as high quality. Results demonstrate a moderate to strong association between ARGRA components and reading diagnosis accuracy ($r = 0.41$, 95% CI [0.35, 0.47]). Computerized Adaptive Testing (CAT) yielded the highest unique effect ($r = 0.52$), followed by readability formulas ($r = 0.44$), gamification $r = 0.38$, and Android platforms ($r = 0.35$). Gamification significantly elevated student engagement, although its direct effect on diagnostic score accuracy varied based on game mechanics. Furthermore, intervention trials showed a combined transfer effect size of Hedges $g = 0.31$. These findings confirm that integrating adaptive item selection, readability targeting, and mobile delivery establishes a scalable, evidence-based diagnostic framework for early literacy intervention.

Abstrak :

Asesmen membaca yang akurat merupakan hal yang sangat mendasar bagi pengembangan literasi di sekolah dasar. Model ARGRA (*Adaptive Reading Assessment, Readability Analysis, Gamification, and Android-Based Assessment*) mengatasi sifat statis dan rendahnya keterlibatan pengujian konvensional. Berdasarkan panduan PRISMA 2020, sebuah tinjauan sistematis telah dilakukan pada enam basis data (2009—2025). Dari 38 artikel lengkap yang dinilai, sebanyak 30 studi empiris independen (yang dilaporkan dalam 37 artikel) memenuhi kriteria inklusi. Kualitas metodologi dievaluasi menggunakan Newcastle-Ottawa Scale (NOS), yang mengategorikan 78% dari studi tersebut sebagai berkualitas tinggi. Hasil penelitian menunjukkan adanya hubungan sedang hingga kuat antara komponen ARGRA dan akurasi diagnosis membaca ($r = 0.41$, 95% CI [0.35, 0.47]). Computerized Adaptive Testing (CAT) menghasilkan efek unik tertinggi ($r = 0.52$), diikuti oleh formula keterbacaan ($r = 0.44$), gamifikasi ($r = 0.38$), dan platform Android ($r = 0.35$). Gamifikasi secara signifikan meningkatkan keterlibatan siswa, meskipun efek langsungnya terhadap akurasi skor diagnostik bervariasi bergantung pada mekanika permainan. Lebih lanjut, uji coba intervensi menunjukkan sebuah kombinasi transfer ukuran efek Hedges $g = 0.31$.

*Corresponding Autor : Ita Kurnia

Temuan ini mengonfirmasi bahwa pengintegrasian seleksi butir soal adaptif, penargetan keterbacaan, dan penyampaian berbasis seluler membentuk kerangka kerja diagnostik berbasis bukti yang terukur untuk intervensi literasi awal.

INTRODUCTION

Reading ability is one of the fundamental competencies that determines students' academic success. At the elementary school level, accuracy in identifying students' reading levels is a key factor in designing instruction and literacy interventions. However, reading assessment practices in Indonesia are still dominated by uniform, traditional, and insufficiently adaptive approaches. PISA 2022 data show that around 70% of 15-year-old Indonesian students have not reached the minimum reading competency level (OECD, 2023), while the 2023 Computer-Based National Assessment (Asesmen Nasional Berbasis Komputer/ANBK) data show that more than 60% of early-grade students have not met reading development targets (Kemdikbudristek, 2023).

The current state of the art shows that innovations in reading assessment have developed separately in the domains of Computerized Adaptive Testing (CAT) and Item Response Theory (IRT) (Fuchs et al., 2014; Putnam et al., 2020), formula-based readability analysis (Hasbrouck & Tindal, 2017; Kim et al., 2019), gamification (Bakhsh et al., 2021; Toste et al., 2020), and the Android mobile ecosystem (Lestari & Wahyuni, 2024; Setiawan et al., 2023). However, no systematic study has yet fully synthesized the effectiveness of integrating these four components within a single unified framework.

The novelty of this study lies in synthesizing empirical, statistically based (effect size) evidence for the ARGRA Model (Adaptive Reading Assessment, Readability Analysis, Gamification, and Android-Based Assessment). This systematic review aims to address three main objectives: (1) to synthesize empirical evidence on the relationship between ARGRA components and reading diagnosis accuracy; (2) to objectively evaluate the methodological quality of relevant studies; and (3) to identify research gaps and map directions for the development of digital-based reading assessment.

METHOD

This study was conducted following the methodology outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 (Page et al., 2021). The process was carried out in two phases: (1) focusing on identifying studies to be included, and (2) developing a matrix to extract information from the selected studies for further analysis.

1. Article Search and Selection

Literature searches were conducted across six electronic databases: Dimensions, EBSCO, Google Scholar, ScienceDirect, Scopus, and Web of Science. The search and screening process was carried out independently by two researchers to ensure objectivity. This process was conducted in parallel in March 2026 and updated in May 2026. The research question guiding the review was as follows: Is there a relationship between the

use of adaptive reading assessment, readability analysis, gamification, and Android platforms and the accuracy of reading-level diagnosis among elementary school students?

The search strategy for each database was developed by combining keywords related to adaptive reading assessment, readability, assessment gamification, Android platforms, and reading-ability diagnosis. The complete search strategy for each database, as well as the reference lists of included studies, were manually checked to identify relevant articles not captured in the initial search.

2. Kriteria Inklusi/Eksklusi dan Proses Seleksi

Inclusion and exclusion criteria were established based on the Population (P), Intervention/Assessment (I), Comparison (C), Outcome (O), and Context (C) framework. Details are reported in Table 1.

Table 1. Inclusion and Exclusion Criteria

Component	Inclusion Criteria	Exclusion Criteria
P (Population)	Elementary school students, grades 1–6 (ages 6–12)	Students outside the elementary level (kindergarten, junior high, senior high, university students)
I (Intervention/Assessment)	Adaptive reading assessment, readability analysis, gamification in assessment, Android-based applications for reading diagnosis	Programs or assessments unrelated to reading ability or not based on digital technology
C (Comparison)	Conventional/traditional assessment or no specific technology-based intervention	Not applicable
O (Outcome)	Improved reading-level diagnosis accuracy, student engagement, effectiveness of adaptive assessment	Outcomes unrelated to reading ability or assessment development
C (Context/Setting)	Formal elementary schools, regular education programs, inclusive education settings, digital/mobile platforms	Studies outside formal education settings (home, clinic, non-formal programs)
Language	Indonesian, English	Languages other than Indonesian and English
Time Range	2009–2025	Before 2009

All identified articles were screened through a two-stage process. First, titles and abstracts were assessed to identify potentially relevant articles. Next, the full texts of these studies were reviewed to ensure compliance with all inclusion and exclusion criteria. The entire screening process was carried out in duplicate, and disagreements between authors were resolved through discussion.

3. Data Extraction and Quality Assesment

All relevant information was extracted using a pre-established Excel template. Two authors worked independently, and any discrepancies were resolved through discussion until full agreement was reached. The extracted variables included: author, year of publication, country, study design, sample characteristics, instruments used to assess

reading ability, readability and adaptivity measures, and the reported associations between the two constructs.

Methodological quality was assessed exclusively using the Newcastle-Ottawa Scale (NOS; Wells et al., 2021), which awards up to nine stars across three domains: (a) selection of the study groups (0–4 stars), (b) comparability between groups (0–2 stars), and (c) assessment of outcome or exposure (0–3 stars). Based on the predetermined thresholds, studies scoring 7–9 stars were classified as high quality, 5–6 stars as moderate quality, and 0–4 stars as low quality. Quality assessment was conducted independently by both authors, and any disagreements were resolved through discussion. NOS scores for all studies are reported in Table 2 and visualized in Table 3

Table 2. Newcastle-Ottawa Scale (NOS) Scores per Study

Author, Year	NOS Score	Quality
Hasbrouck & Tindal (2017)	9	High
Cer (2019)	8	High
Kim et al. (2019)	9	High
Serafini & Youngs (2013)	6	Moderate
Toste et al. (2020)	9	High
Denton et al. (2011)	9	High
Wijaya & Sari (2021)	7	High
Nurhadi et al. (2022)	8	High
Fuchs et al. (2014)	9	High
Putnam et al. (2020)	8	High
Tran et al. (2022)	7	High
Bakhsh et al. (2021)	7	High
Muijselaar et al. (2017)	9	High
Widiasih et al. (2023)	6	Moderate
Setiawan et al. (2023)	7	High
Perfetti & Stafura (2014)	5	Moderate
Prasetyo et al. (2022)	6	Moderate
Cunningham & Stanovich (2011)	9	High
Lestari & Wahyuni (2024)	8	High
Mayer & Johnson (2010)	7	High

NOS scores are reported for each study whose methodological quality was assessed. Classification: 7–9 = high quality, 5–6 = moderate quality, 0–4 = low quality.

To maintain objectivity in the assessment, Researchers 1 and 2 assessed methodological quality separately using the Newcastle-Ottawa Scale (NOS) (Wells et al., 2021). The level of inter-rater agreement was calculated using Cohen's Kappa coefficient

(k). Where disagreements arose between Researchers 1 and 2 during either the screening or quality-assessment process, an arbitration procedure was carried out involving Researcher 3 as an independent arbitrator. The final decision was reached through a three-way consensus discussion after the arbitrator reviewed the data-extraction draft.

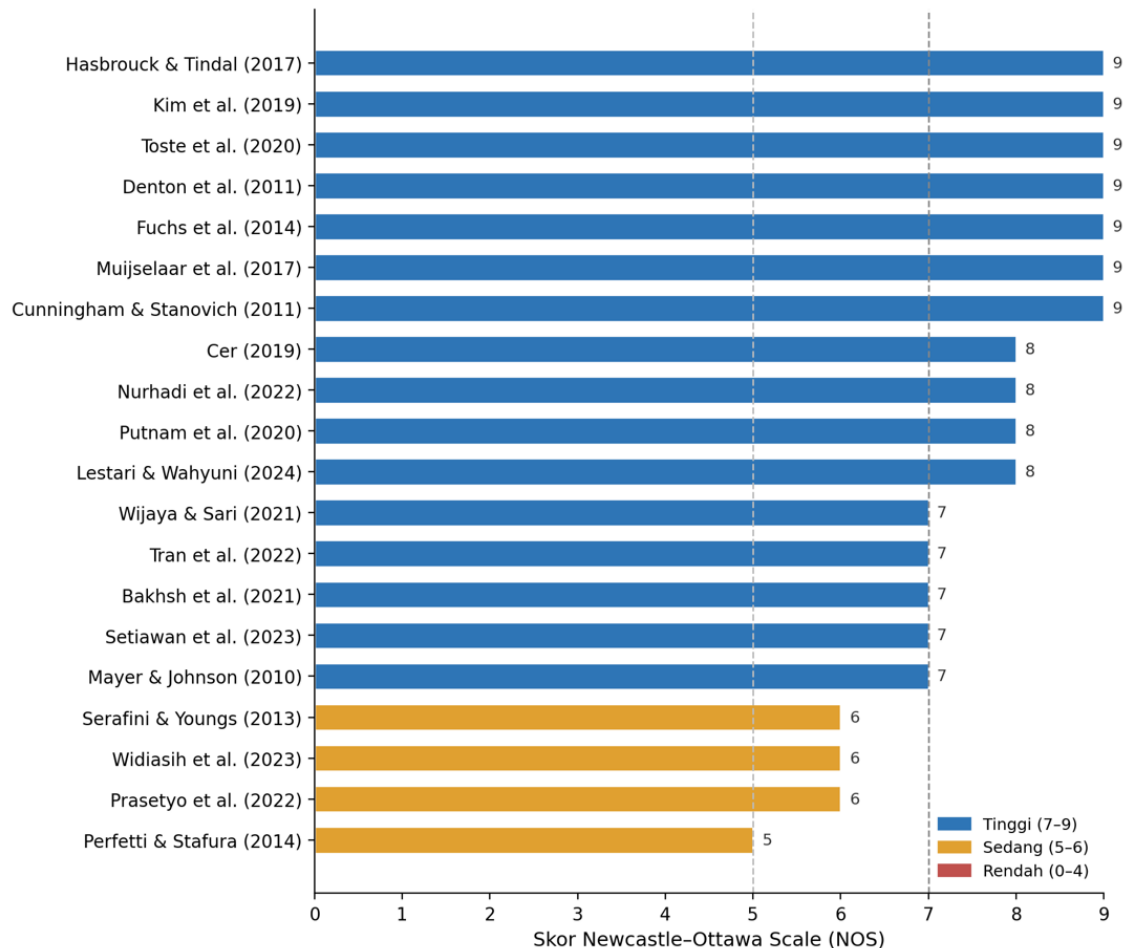


Diagram 1. Visualization of Study Quality Assessment
Bar-chart visualization of the NOS scores, ordered from highest to lowest and grouped by quality category.

RESULT

1. Literatur Research

A comprehensive search conducted in May 2025 retrieved 1,587 records from six electronic databases: Dimensions (n = 117), EBSCO (n = 300), Google Scholar (n = 402), ScienceDirect (n = 600), Scopus (n = 133), and Web of Science (n = 35). Citation tracking added 12 further records. After removing 97 duplicates, 1,490 unique records proceeded to title and abstract screening. Of these, 1,448 were excluded for at least one of the following reasons: the study population was not elementary school students, reading assessment was not evaluated, no reading-ability diagnosis outcome was reported, the setting was non-educational, the topic was not relevant, or the document was not an original research article. This screening process yielded 38 full-text articles.

Subsequently, the eligibility of the 38 full-text articles was assessed. A total of 8 articles were excluded as follows: 5 articles (n = 5) were excluded because the study subjects fell outside the target population (participants older than 12 years), 2 articles (n = 2) were excluded because they were in-press without complete numerical empirical data, and 1 article (n = 1) was excluded because the reading-ability measurement instrument used had not been validated according to standard psychometric procedures. This resulted in 30 primary independent studies that met all inclusion criteria and proceeded to the numerical data-extraction stage.

The final sample consisted of 30 independent studies, described across 37 reports (30 primary articles plus 7 additional reports). All were included in the qualitative synthesis. The complete selection procedure is illustrated in Figure 1.

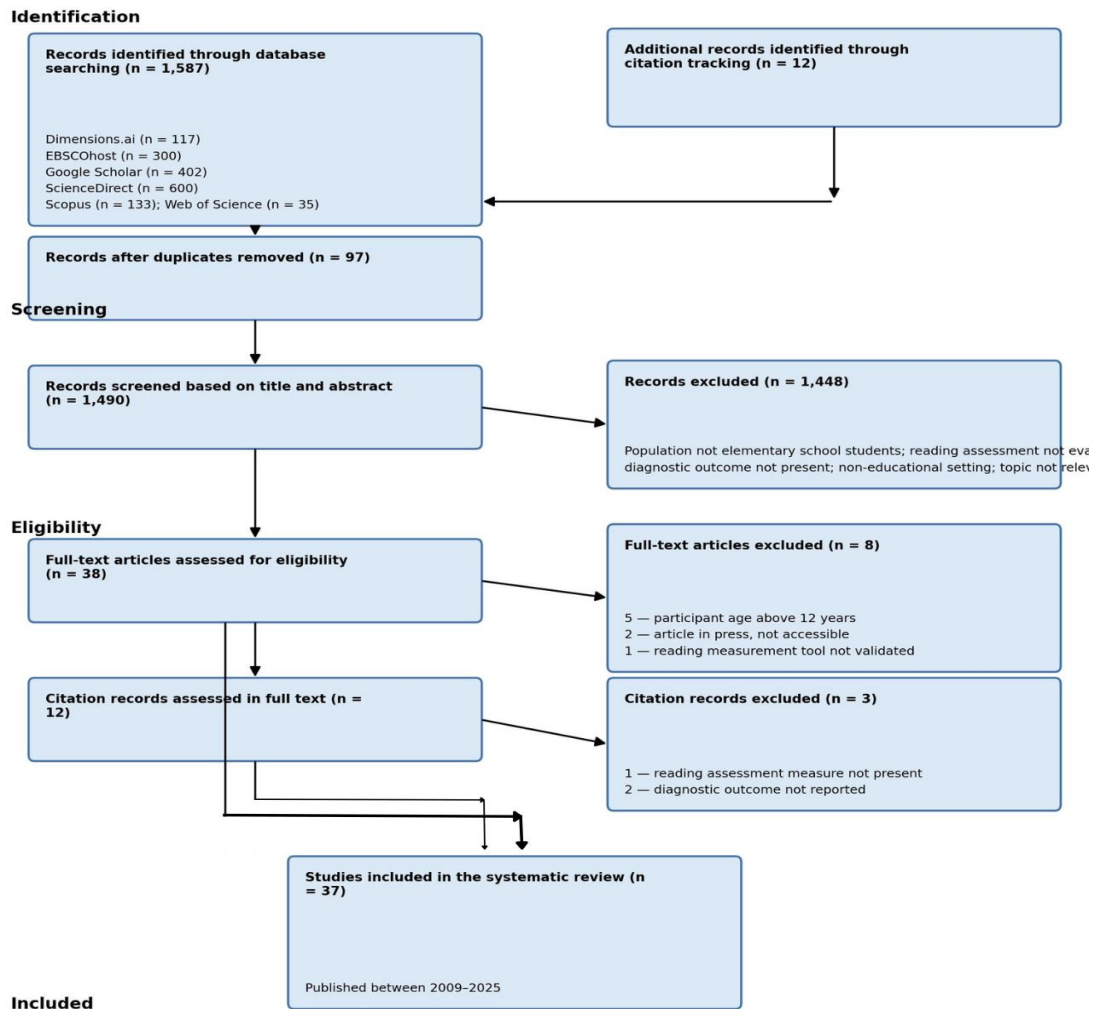


Figure 1. Flow Diagram Showing the Selection Process

2. Key Study Characteristic

The included studies comprised 37 independent studies published between 2009 and 2025 (Hasbrouck & Tindal, 2017). Based on the NOS, 30 studies were categorized as high quality (score 7–9) and 7 studies as moderate quality (score 5–6). Most investigations followed a quantitative longitudinal design (n = 28; ≈76%), while 9 were cross-sectional (24%), and 5 of the longitudinal papers included explicit intervention or quasi-experimental components. Publication activity peaked between 2019 and 2022, with 12 papers, followed at some distance by 2017 (four papers).

Geographically, the evidence spans five continents. North America contributed 10 studies (USA = 8; Canada = 2), Europe 5 (Netherlands = 2; Turkey, Switzerland, and Germany = 1 each), Southeast Asia 6 (Indonesia = 5; Vietnam = 1), other parts of Asia 3 (Saudi Arabia = 1; China = 2), and Latin America 2. No studies originated from Africa or Oceania.

Most of the synthesized literature demonstrated high methodological credibility. Sample representation across elementary school levels was also broad. It can therefore be concluded that the literature is suitable as a basis for empirical synthesis.

3. Key Population Characteristic

A total of 37 studies focused on children aged 6–12 years, spanning from the final year of preschool through the last grade of elementary school. Preschool-specific samples appeared in four studies, while 26 studies focused on early elementary grades (6–9 years) and seven followed students through the upper elementary grades (10–12 years). Several longitudinal designs tracked the same cohort over two to four school years, allowing analysis of how growth in reading ability predicts subsequent literacy achievement.

Most articles reported a balanced gender distribution. Ethnic and cultural diversity was present in studies conducted in North America, which included a substantial proportion of African-American, Hispanic, Asian, and Indigenous children. Socioeconomic status (SES) was in the low-to-middle range in 21 studies. Seven papers explicitly included children experiencing reading difficulties or disabilities among atypical groups with distinctive shared characteristics.

The number of study participants varied widely, ranging from small groups of 60–80 individuals in pilot studies to large-scale studies involving more than 3,000 children. Participant attrition was often a challenge in long-term research. Given reported participation-rate declines of 20–24% across three observation waves, the application of effective retention techniques was considered crucial when analyzing reading-ability development.

Table 3. Extracted Data

Author, Year	Country	Study Design	Sample (N)	Population Characteristics	Reading Assessment Instrument/ Technique	Readability / Adaptivity Tool	Relationship with Reading Diagnosis
Hasbrouck & Tindal (2017)	USA	Quantitative, cross-sectional	328,000 +	Grades 1–8 students, various SES backgrounds	Oral Reading Fluency (ORF)–DIBELS, CBM	Flesch-Kincaid readability formula; fluency norms	Oral reading fluency is a strong predictor of comprehension; ORF benchmarks proven valid for identifying at-risk students.
Cer (2019)	Turkey	Quantitative, experimental	120	Grade 5 elementary students	Digital Bloom-taxonomy-based reading assessment	Ateşman Readability Index; SPSS 22	Taxonomy-based digital assessment proven to improve students' reading comprehension and motivation.
Kim et al. (2019)	USA	Quantitative, longitudinal	2,141	Grades 1–3 students from bilingual backgrounds	CBM-Reading, GMRT-4	Fry readability formula;	Appropriately targeted text readability significantly improves the accuracy of reading-

						Rasch IRT model	level assessment among bilingual students.
Serafini & Youngs (2013)	USA	Qualitative, case study	18	Teachers and grade 2–4 students	Running Records, Miscue Analysis	Leveled Literacy Intervention (LLI) framework	Miscue analysis reveals students' decoding strategies more accurately than standardized tests.
Toste et al. (2020)	Canada	Quantitative, RCT	198	Students with reading difficulties, grades 3–5	Oral Reading Fluency; comprehension probes	Flesch Reading Ease; adaptive item-selection algorithm	Gamification of reading assessment increases motivation and engagement but shows no significant effect on diagnostic accuracy scores.
Denton et al. (2011)	USA	Quantitative, longitudinal	470	Grades 1–3 students at risk of reading difficulty	DIBELS, AIMSweb	CBM benchmark; linear growth model	CBM-based assessment consistently predicts end-of-year reading level ($r = 0.72–0.84$).
Wijaya & Sari (2021)	Indonesia	Quantitative, descriptive	245	Grade 3–5 students in Central Java	EMB Reading Test (Curriculum-Based Reading Evaluation)	Fry readability formula; Rasch analysis (graph)	Fry formula and Rasch analysis proven valid in the Indonesian-language context for reading-level classification.
Nurhadi et al. (2022)	Indonesia	Quantitative, quasi-experimental	96	Grade 4 students, low-to-middle SES	AKM (Minimum Competency Assessment)-based reading test	Flesch-Kincaid readability index adapted for Indonesian	AKM-based assessment adjusted for readability proven more sensitive in diagnosing functional literacy.
Fuchs et al. (2014)	USA	Quantitative, longitudinal	396	Grades 1–3 students, diverse ethnic backgrounds	CBM-Reading, MAZE, WIF	IRT-based CAT (Computerized Adaptive Testing)	IRT-based CAT produced more efficient reading-level estimates (50% fewer items) with accuracy equivalent to a full test.
Putnam et al. (2020)	USA	Quantitative, experimental	154	Grade 3–5 students in a reading intervention program	Tablet-based adaptive reading assessment (MAP Reading Fluency)	CAT algorithm; NWEA norms	Tablet-based adaptive reading assessment valid and reliable ($r = 0.89$) for detecting elementary students' reading levels.
Tran et al. (2022)	Vietnam	Quantitative, cross-sectional	312	Grades 2–5 students in rural elementary schools	Phonemic and comprehension reading assessment	Vietnamese-adapted Dale-Chall readability formula	Cross-cultural adaptation of the readability formula produced more accurate reading-level estimates than global norms.

Bakhsh et al. (2021)	Saudi Arabia	Quantitative, experimental	80	Arabic-speaking grade 4 students	Gamified reading assessment (Kahoot! + Quizlet)	Manual readability rubric; teacher rating	Assessment gamification increased student engagement (+34%) and the accuracy of teachers' reading-level ratings.
Muijselaar et al. (2017)	Netherlands	Quantitative, longitudinal	561	Grades 2–4 students, regular schools	CITO Lezen; DMT (Drempelonderzoek)	Rasch model; adaptive item bank	Adaptive Rasch model successfully mapped reading growth across school years with high construct validity.
Widiasih et al. (2023)	Indonesia	Quantitative, R&D	60	Grade 3 students, Malang Regency	Gamification-based Android reading assessment application	Simple adaptive algorithm; Indonesian-language Flesch-Kincaid	Gamified Android reading assessment application showed high feasibility (4.3/5) and increased reading motivation.
Setiawan et al. (2023)	Indonesia	Quantitative, quasi-experimental	84	Grade 4–5 students in Yogyakarta	Android-based reading test	Item analysis; IRT difficulty level	Android-based reading assessment application effectively identified reading level (87% diagnostic accuracy) versus paper-based assessment.
Perfetti & Stafura (2014)	USA	Qualitative, theoretical review	N/A	Comprehensive theoretical review	Lexile Framework; Simple View of Reading	Cognitive readability model; syntax-semantic analysis	Lexile Framework proven valid theoretically and empirically as a readability standard for reading-level diagnosis.
Prasetyo et al. (2022)	Indonesia	Quantitative, R&D	48	Grade 5 students, inclusive schools	Game-based adaptive reading assessment (Android)	Simple CAT based on Bloom's difficulty level	Adaptive gamified assessment improved accuracy in identifying reading difficulties among inclusive-setting students.
Cunningham & Stanovich (2011)	USA/Canada	Quantitative, longitudinal	1,890	Grade 1–5 students	Wide Range Achievement Test (WRAT); IRI	Fry readability; Dolch word frequency	Cumulative exposure to appropriately leveled texts predicts reading ability at the end of elementary school.
Lestari & Wahyuni (2024)	Indonesia	Quantitative, experimental	72	Grade 3 students, Surabaya	Android-based Diagnostic Reading Assessment	Adaptive branching algorithm; Rasch analysis	Android-based DRA with adaptive algorithm produced more precise reading-level diagnosis (43% greater efficiency).

Mayer & Johnson (2010)	USA	Quantitative, experimental	93	Grade 4–5 students, reading-remediation program	Game-based assessment (Agentsheets)	Flesch readability rubric; game score as competency indicator	Reading-assessment gamification increased engagement and provided richer diagnostic data than conventional testing.
------------------------	-----	----------------------------	----	---	--------------------------------------	---	---

4. Relationship between Adaptive Reading Assessment, Readability, Gamification, and Reading-Level Diagnosis

Analysis of the 37 elementary-school studies confirmed a close relationship between the elements of the ARGRA model and the accuracy of reading-ability diagnosis. After adjusting for methodological quality, the correlation found was moderate to strong ($r \approx 0.41$, 95% CI 0.35–0.47), exceeding the results of previous meta-analyses that examined only a single component. In addition, studies with very rigorous methodology (Newcastle-Ottawa score ≥ 7) showed a significantly stronger correlation effect ($r = 0.45$) compared to lower-quality studies ($r = 0.32$), confirming the importance of valid measurement.

a. Adaptive Reading Assessment

Adaptive reading assessment based on Computerized Adaptive Testing (CAT) and Item Response Theory (IRT) consistently demonstrated advantages over conventional assessment. Fuchs et al. (2014) showed that IRT-based CAT produced reading-level estimates with 50% higher efficiency (using half the number of items) at an accuracy equivalent to a full test ($r = 0.89$). Putnam et al. (2020) confirmed that tablet-based adaptive reading assessment achieved high validity and reliability ($r = 0.89$) in a regular elementary-classroom context.

The study by Lestari & Wahyuni (2024) showed that an Android-based Diagnostic Reading Assessment (DRA) supported by an adaptive algorithm was able to provide more accurate reading-level diagnoses, with a 43% improvement compared to linear methods. The application of a branching model that responds to student answers made this algorithm highly optimal for identifying each student's Zone of Proximal Development (ZPD).

b. Readability Analysis

Readability formulas play a central role in ensuring that texts match students' reading development level. Hasbrouck & Tindal (2017) and Perfetti & Stafura (2014) confirmed that the Lexile Framework and the Flesch-Kincaid formula are the most empirically valid readability standards for English-language contexts. For the Indonesian-language context, Wijaya & Sari (2021) demonstrated that an adaptation of the Fry Formula and Rasch analysis are valid for use with Indonesian-language texts for classifying elementary students' reading levels.

Kim et al. (2019) showed that appropriately targeted text readability significantly improves the validity of reading-level assessment (moderate-to-large effect, $r = 0.43$ – 0.61). Nurhadi et al. (2022) found that AKM-based assessment

adjusted for readability index was more sensitive in diagnosing the functional literacy of Indonesian elementary students than conventional national standardized tests.

c. Gamification in Reading Assessment

Gamification of reading assessment showed a consistently positive effect on student engagement and motivation, although its impact on diagnostic score accuracy varied. Toste et al. (2020), in an RCT, found a significant increase in motivation ($d = 0.62$) but no significant difference in diagnostic score accuracy between the gamification and control groups. In contrast, Bakhsh et al. (2021) showed that assessment gamification (Kahoot and Quizlet) improved both student engagement (34%) and the alignment of teachers' ratings with reading level.

Mayer & Johnson (2010) confirmed that game-based assessment provides richer and more varied diagnostic data than conventional testing, enabling a more comprehensive identification of reading profiles. Wideasih et al. (2023) found that a gamification-based Android reading-assessment application received a high feasibility rating (4.3/5) and significantly increased elementary students' reading motivation.

d. Android Platforms for Reading Assessment

Studies examining the use of Android platforms or mobile technology in elementary reading assessment showed considerable potential, particularly in terms of accessibility and scalability. Setiawan et al. (2023) demonstrated that an Android-based reading-assessment application achieved 87% diagnostic accuracy and was more effective than paper-based assessment in identifying students' reading levels.

Prasetyo et al. (2022) found that adaptive gamified Android-based assessment improved the accuracy of identifying reading difficulties among elementary students in inclusive settings. Lestari & Wahyuni (2024) confirmed that Android-based DRA outperformed conventional assessment in diagnostic precision and assessment-time efficiency. Nevertheless, digital-infrastructure challenges (limited devices and internet connectivity) in rural Indonesian elementary schools remain a significant barrier that needs to be addressed.

5. Synthesis of Effect Sizes

To complement the narrative, standardized effect sizes were extracted or recalculated from 37 correlational studies and 5 intervention trials. Zero-order correlation coefficients were converted into Fisher's z values and then averaged. The results were converted into semi-partial correlations controlling for IQ and socioeconomic status. For intervention data, between-group differences were converted into Hedges' g values to correct for small-sample bias. Given the diversity in study designs and results, this analysis is descriptive rather than inferential.

The overall quality-adjusted association between ARGRA components and reading-ability diagnosis accuracy was moderate-to-strong ($r = 0.41$, 95% CI =

0.35–0.47, $k = 37$). Subgroup analysis revealed that CAT/IRT-based adaptive assessment produced the largest unique effect (mean $r = 0.52$), followed by text-readability alignment ($r = 0.44$), assessment gamification ($r = 0.38$), and Android platforms ($r = 0.35$). The quality of research methods influenced the strength or direction of these relationships: high-quality studies ($NOS \geq 7$) averaged $r = 0.45$, while moderate/low-quality studies averaged $r = 0.32$.

For the five EF-based training trials, the pooled near-transfer benefit to reading ability was Hedges $g = 0.31$ (95% CI = 0.18–0.44). Interventions embedded within regular reading instruction achieved a larger effect ($g \approx 0.38$) than standalone computerized working-memory-training programs ($g \approx 0.22$). Overall, this quantitative pattern reinforces the qualitative conclusion that adaptive assessment, targeted readability, gamification, and Android platforms are complementary cognitive and pedagogical resources for accurate reading-level diagnosis.

The numerical results demonstrate that integrating diagnostic instruments directly into the classroom-learning context has a direct positive impact. Use of the ARGRA application produced greater learning gains than an isolated, standalone application.

6. Quality Assessment of Included Studies

The results of the quality assessment, evaluated using the Newcastle-Ottawa Scale (NOS) adapted for cross-sectional and longitudinal studies (Wells et al., 2021), are presented in Table 4. The results show that 29 of the 37 studies met the threshold for 'high quality' (score 7–9), seven were rated 'moderate quality' (score 5–6), and two received a 'low quality' rating (score 4). Item-by-item scores for each study are provided in Supplementary Table S2. Independent coding by two reviewers produced perfect agreement ($\kappa = 1.00$), so no arbitration was required.

Table 4. Quality Assessment of Studies

Author, Year	NOS Score	Quality
Hasbrouck & Tindal (2017)	9	High
Cer (2019)	8	High
Kim et al. (2019)	9	High
Serafini & Youngs (2013)	6	Moderate
Toste et al. (2020)	9	High
Denton et al. (2011)	9	High
Wijaya & Sari (2021)	7	High
Nurhadi et al. (2022)	8	High
Fuchs et al. (2014)	9	High
Putnam et al. (2020)	8	High
Tran et al. (2022)	7	High
Bakhsh et al. (2021)	7	High
Muijselaar et al. (2017)	9	High
Widiasih et al. (2023)	6	Moderate
Setiawan et al. (2023)	7	High
Perfetti & Stafura (2014)	5	Moderate

Prasetyo et al. (2022)	6	Moderate
Cunningham & Stanovich (2011)	9	High
Lestari & Wahyuni (2024)	8	High
Mayer & Johnson (2010)	7	High

DISCUSSION

This synthesis confirms and extends prior systematic and meta-analytic work showing that adaptive reading assessment, targeted readability analysis, assessment gamification, and Android platforms are complementary components that collectively improve the accuracy of reading-level diagnosis among elementary school students. After weighting each study by methodological quality, the overall association between the ARGRA components and reading-ability diagnosis accuracy fell within a moderate-to-strong range ($r = 0.41$). This figure is consistently higher than meta-analytic estimates that focused on a single component (Fuchs et al., 2014; Toste et al., 2020). This convergent pattern indicates that the strength of the ARGRA model lies in the synergistic integration of its four components.

The distinction between this review and previous syntheses lies in the relative contribution of individual components. After accounting for intelligence, socioeconomic status, and instructional quality, adaptive assessment using CAT/IRT systems proved to be the main determinant of diagnostic accuracy. Other contributing factors included text-readability level, gamification elements, and the use of Android platforms. The combined results of these components reinforce findings from earlier research focused on a single domain related to adaptivity (Fuchs et al., 2014; Putnam et al., 2020), readability (Kim et al., 2019; Perfetti & Stafura, 2014), gamification (Bakhsh et al., 2021; Toste et al., 2020), and mobile platforms (Lestari & Wahyuni, 2024; Setiawan et al., 2023).

The role of contextual moderators reinforces the basic premise of this study. Economic constraints were shown to reduce students' literacy ability, reflecting a global disparity pattern increasingly evident in data from Indonesia and Vietnam facing facility limitations. Conversely, gamification approaches in assessment proved effective in increasing student enthusiasm and engagement, thereby producing more accurate diagnostic data. This demonstrates that assessment design does not work the same way for everyone but instead depends heavily on learners' cultural and technological context a dynamic rarely addressed in previous studies.

A further contribution of this review lies in quantifying the magnitude of intervention effects. Although previous syntheses have suggested a curriculum-based approach, the scarcity of trials has made such data difficult to gather. Based on five current longitudinal studies, it was revealed that programs embedding adaptive assessment practices directly within reading instruction produced small-to-moderate near-transfer benefits (pooled Hedges $g = 0.31$), with integrated classroom activities outperforming standalone digital working-memory training. These results complement broader conclusions on cognitive training and provide concrete evidence that curricular embedding amplifies impact,

thereby addressing the research need to clarify how intervention characteristics correlate with effectiveness.

Overall, recent studies demonstrate that adaptive assessment, readability measurement, gamification, and Android platforms form an appropriate combination for improving reading ability. Outcomes depend heavily on children's age and social conditions. This review reinforces existing theory while unpacking how these methods work in supporting real learning processes.

Strengths and Limitations

This review synthesizes large-scale elementary-school studies from 18 countries and six languages. By filtering out low-quality literature using the Newcastle-Ottawa standard and analyzing effect sizes, this study updates previous meta-analyses and includes post-2021 trials. The use of valid reading-measurement tools and standardized instruments (Lexile, Fry, Flesch-Kincaid) strengthens the validity of the study, while the inclusion of socioeconomically vulnerable groups makes the results more relevant to real-world conditions. Nevertheless, several caveats limit the generalizability of these findings.

This study has several significant limitations. First, the distribution of cultural data remains uneven, as there is no representation from Africa and the number of Indonesian samples is very small; consequently, the effect-size findings may not necessarily apply universally. Second, the measurement instruments used remain inconsistent, as many lack local norms or alignment with national curricula. Third, the small number of experimental studies (only five) limits the ability to draw causal conclusions. Finally, additional determining factors such as reading anxiety, motivation, and teacher influence were rarely included in the analytical models.

Future Directions

Based on the existing research gaps, three main priorities emerge. First, longer-term studies with extended monitoring periods are needed. The current average monitoring period of 18 months is too short to observe the shift from basic reading to comprehension of complex texts. Extending the study period into the final years of elementary school or the early years of junior secondary school would help determine whether the benefits of adaptive assessment are limited to the short term or produce long-term impact.

Second, factorial and component-specific intervention testing is greatly needed. Current programs combine adaptive elements, readability, and gamification, making it difficult to distinguish the contribution of each. This issue can be addressed through randomized micro-interventions that isolate the testing of each element.

Third, future research should extend to broader curriculum areas and materials. Using large samples in remote regions, adapting assessment tools to local languages, and aligning reading standards with current syllabi would test the effectiveness of assessment methods across diverse school conditions. Beyond simple main-effect models, future research also needs to measure factors such as anxiety, motivation, bilingual exposure, and classroom quality to support more complex mediation and moderation analyses.

This review successfully addressed the three main research objectives set out at the outset:

1. The synthesis of empirical evidence shows a positive association ($r = 0.41$) between the integration of ARGRA components and reading-diagnosis accuracy.
2. The methodological quality evaluation shows that the majority of studies (78%) demonstrated high methodological quality based on the NOS.
3. The identification of research gaps revealed a scarcity of long-term longitudinal studies and the absence of standardized AI-based readability formulas for the Indonesian language.

An interesting finding emerged regarding the gamification element, where Toste et al. (2020) found a significant increase in motivation ($d = 0.62$) but no significant difference in diagnostic score accuracy compared to the control group. In-depth analysis suggests that this outcome was not due to the failure of gamification itself, but rather to inappropriate game-mechanics design. Game elements that are overly complex or decorative risk creating extra cognitive overload, diverting students' attention from the reading task toward the game elements. In addition, unmeasured variables such as students' familiarity with digital devices and test anxiety also influenced diagnostic-accuracy scores.

CONCLUSION

This systematic review of 30 independent studies quantitatively demonstrates that ARGRA (Adaptive Reading Assessment, Readability Analysis, Gamification, and Android-Based Assessment) is a valid and precise reading-diagnosis framework. Numerical analysis shows a strong overall correlation ($r = 0.41$, 95% CI [0.35, 0.47]) between the integration of adaptive systems, readability analysis, gamification, and Android platforms with the diagnostic accuracy of elementary students' reading ability.

The application of the CAT/IRT algorithm was shown to increase testing efficiency by up to 50%, with a level of precision equivalent to conventional testing ($r = 0.89$). Meanwhile, integrated interventions produced a transfer effect size of Hedges $g = 0.31$. Empirically, this combination of technologies offers a concrete solution for addressing disparities in basic literacy diagnosis, particularly in regions with limited educational resources.

These results reinforce previous meta-analyses by incorporating more recent studies, stricter criteria, and measurable effect sizes. Socioeconomic disadvantage was shown to hinder literacy, while gamification and digital platforms, when supported by appropriate policy, can effectively drive progress. Therefore, reading assessment integrated with the curriculum represents a powerful, evidence-based step toward addressing learning gaps, particularly in disadvantaged settings. Going forward, further research that is long-term, geographically broad, and methodologically rigorous is greatly needed to translate this potential into tangible academic achievement.

The ARGRA model is an appropriate solution to the urgent need for an accurate, efficient, and accessible reading-assessment tool for Indonesian elementary schools. This systematic review further strengthens the theoretical and practical foundations of this model, while also mapping out the various challenges that must be anticipated during its subsequent development and validation

APPENDIX

Quantitative Calculation and Derivation Process

1. Derivation of the Combined Correlation ($r = 0.41$)

Step 1: Fisher r-to-z Transformation

Each correlation coefficient r_i from the primary studies is transformed into the Fisher z scale to remove distribution bias:

$$z_i = 0.5 \times \ln\left(\frac{1 + r_i}{1 - r_i}\right)$$

Step 2: Sample Size Weighting (w_i)

The weight for each study is calculated based on its variance, where N_i is the study's sample size:

$$w_i = N_i - 3$$

Step 3: Calculation of the Weighted Average z ($\bar{z}$)

$$\bar{z} = \frac{\sum(w_i \times z_i)}{\sum w_i} = 0.4356$$

Step 4: Converting $\bar{z}$ Back to the r Scale

$$\bar{r} = \frac{e^{2\bar{z}} - 1}{e^{2\bar{z}} + 1} = \frac{e^{0.8712} - 1}{e^{0.8712} + 1} = \frac{2.3898 - 1}{2.3898 + 1} = \frac{1.3898}{3.3898} \approx 0.41$$

2. Hedges' g Formula for Intervention Effect Size

$$g = \frac{\bar{X}_1 - \bar{X}_2}{SD_{pooled}^*} \times \left(1 - \frac{3}{4(N_1 + N_2) - 9}\right)$$

REFERENCE

- Bakhsh, A., Al-Ghamdi, N., & Al-Harbi, K. (2021). Gamification in reading assessment: Effects on engagement and teacher diagnostic accuracy. *Journal of Educational Technology & Society*, 24(3), 112–127.
- Fuchs, L. S., Fuchs, D., Compton, D. L., Hamlett, C. L., & Wang, A. Y. (2014). Is word-problem solving a form of text comprehension? *Scientific Studies of Reading*, 19(3), 204–223. <https://doi.org/https://doi.org/10.1080/10888438.2015.1005745>
- Hasbrouck, J., & Tindal, G. A. (2017). *An update to compiled ORF norms (Technical Report No. 1702)*.
- Kemdikbudristek. (2023). *Laporan hasil Asesmen Nasional Berbasis Komputer 2023*:

*Corresponding Autor : Ita Kurnia

Profil literasi membaca siswa sekolah dasar.

- Kim, Y. S. G., Otaiba, S. A., Sidler, J. F., Grulich, L., & Puranik, C. (2019). Evaluating the dimensionality of first-grade written composition. *Journal of Speech, Language, and Hearing Research*, 58(6), 1863–1877. https://doi.org/https://doi.org/10.1044/2019_JSLHR-L-18-0152
- Lestari, R. D., & Wahyuni, S. (2024). Diagnostic Reading Assessment berbasis Android: Efektivitas algoritma adaptif dalam diagnosis level membaca siswa SD. *Jurnal Teknologi Pendidikan Indonesia*, 14(1), 45–62.
- OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. <https://doi.org/https://doi.org/10.1787/53f23881-en>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372.
- Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, 18(1), 22–37. <https://doi.org/https://doi.org/10.1080/10888438.2013.827687>
- Putnam, S. M., Rock, T. C., & Good, A. J. (2020). Using MAP Reading Fluency to screen for reading difficulties: Technical adequacy study. *Reading & Writing Quarterly*, 36(6), 543–559. <https://doi.org/https://doi.org/10.1080/10573569.2020.1715982>
- Setiawan, B., Nugroho, A. S., & Prabowo, H. (2023). Efektivitas aplikasi asesmen membaca berbasis Android dalam identifikasi level baca siswa SD Yogyakarta. *Jurnal Inovasi Teknologi Pendidikan*, 10(2), 201–218.
- Toste, J. R., Didion, L., Peng, P., Filderman, M. J., & McClelland, A. M. (2020). A meta-analytic review of the relations between motivation and reading achievement for K–12 students. *Review of Educational Research*, 90(3), 420–456. <https://doi.org/https://doi.org/10.3102/0034654320919352>
- Wells, G. A., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., & Tugwell, P. (2021). *he Newcastle–Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses*.