Mathematics Education And Application Journal (META)
Volume 7, No. 2, December 2025, pp. 118-135

DEVELOPMENT OF A NUMERACY LITERACY TEST INSTRUMENT BASED ON THE ANDROID ISPRING SUITE FOR GRADE VIII JUNIOR HIGH SCHOOL STUDENTS

Elwan Stiadi[1]*, Yusnia[2], Nader U-sin[3]
[1]Universitas Bengkulu, Indonesia
[2]Universitas Bengkulu, Indonesia
[3]Principle of Ban Lanchang School, Thailand

*Corresponding author
elwanstiadi@unib.ac.id

Received: 30 October 2025 / Revised: 30 December 2025 / Accepted: 31 December 2025 /
Published online: 31 December 2025

Abstract
The rapid development of technology and communication has penetrated the field of education, including in the development of evaluation instruments. Android-based test instruments have been widely studied and proven to have a positive impact on the learning process. This study aims to develop an Android-based numeracy literacy test instrument using iSpring Suite for Grade VIII junior high school students. This research focused on 30 students in Class VIII B at SMP Negeri 6 Bengkulu City. The results of the development are expected to produce a test instrument that is valid, reliable, practical, and has good item quality. This study employs the ADDIE development model which includes five stages: Analysis, Design, Development, Implementation, and Evaluation. The Analysis stage was conducted to identify the needs of the test instrument, while the Design stage involved the design of an Android-based instrument. At the Development stage, validation was carried out by expert validators using a 4-point Likert scale, which includes the validity of the material, construction, language, and media. The developed product prototype was demonstrated through video documentation, while the application was distributed as an APK file for field testing and user feedback collection. Implementation data was analyzed using the SPSS program to evaluate the reliability and practicality of the instrument. Expert validation yielded an average validity score of 3.82 (Aiken's V = 0.94, very valid category), with a reliability coefficient of r_{11} = 0.890 (high reliability). The practicality test results from students showed a score of 85.2% (very practical category). The results of the study showed that the developed test instrument met the criteria for validity, reliability, and practicality. In addition, the test items have a balanced level of difficulty (2 medium, 2 difficult), good discriminating power (D range: 0.333–0.467), and optimally functioning distractors. This instrument contributes to the development of digital-based evaluation media to improve the quality of numeracy literacy learning aligned with Kurikulum Merdeka and AKM frameworks.

Keywords: Android, ISpring Suite, Numeracy Literacy, Test Instrument.

Abstrak
Perkembangan teknologi dan komunikasi yang pesat telah merambah bidang pendidikan, termasuk dalam pengembangan instrumen evaluasi. Instrumen tes berbasis Android telah banyak dipelajari dan terbukti memiliki dampak positif pada proses pembelajaran. Penelitian ini bertujuan untuk mengembangkan instrumen tes literasi numerasi berbasis Android menggunakan iSpring Suite untuk siswa kelas VIII SMP. Penelitian ini berfokus pada 30 siswa kelas VIII B di SMP Negeri 6 Kota Bengkulu. Hasil pengembangan diharapkan menghasilkan instrumen tes yang valid, reliabel, praktis, dan memiliki kualitas item yang baik. Penelitian ini menggunakan model pengembangan ADDIE yang mencakup lima tahapan: Analisis, Desain, Pengembangan, Implementasi, dan Evaluasi. Tahap Analisis dilakukan untuk mengidentifikasi kebutuhan instrumen tes, sedangkan tahap Desain melibatkan desain instrumen berbasis Android. Pada tahap Pengembangan, validasi dilakukan oleh validator ahli menggunakan skala Likert 4 poin, yang meliputi validitas materi, konstruksi, bahasa, dan media. Prototipe produk yang dikembangkan didemonstrasikan melalui dokumentasi video, sedangkan aplikasi didistribusikan sebagai file APK untuk pengujian lapangan dan pengumpulan umpan balik pengguna. Data implementasi dianalisis menggunakan program SPSS untuk mengevaluasi reliabilitas dan kepraktisan instrumen. Validasi ahli menghasilkan skor validitas rata-rata 3,82 (Aiken's V = 0,94, kategori sangat valid), dengan koefisien reliabilitas r_{11} = 0,890 (reliabilitas tinggi). Hasil uji kepraktisan dari siswa

Copyright © 2025 The Author (s)
This is an open access article under the CC BY SA License

menunjukkan skor 85,2% (kategori sangat praktis). Hasil penelitian menunjukkan bahwa instrumen tes yang dikembangkan memenuhi kriteria validitas, reliabilitas, dan kepraktisan. Selain itu, item tes memiliki tingkat kesulitan yang seimbang (2 sedang, 2 sulit), daya diskriminasi yang baik (rentang D: 0,333–0,467), dan distraktor yang berfungsi optimal. Instrumen ini berkontribusi pada pengembangan media evaluasi berbasis digital untuk meningkatkan kualitas pembelajaran literasi numerik yang selaras dengan kerangka kerja Kurikulum Merdeka dan AKM.

Kata kunci: *Android, Instrumen Tes, ISpring Suite, Literasi Numerasi.*

INTRODUCTION

Mathematics education aims to develop students' critical, logical, and systematic thinking to solve real-world problems efficiently and effectively. However, Indonesia's performance in PISA 2018 revealed that students' numeracy literacy remains low, ranking 74th out of 79 countries with an average score of 379, significantly below the OECD average of 489. This indicates an urgent need to improve numeracy literacy assessment instruments that align with 21st-century demands and facilitate Higher-Order Thinking Skills (HOTS).

In the world of education, assessment serves as a critical component for determining student achievement and informing instructional decisions. The quality of assessment instruments directly affects the accuracy of student achievement results and decisions made by educators and schools. Current 2013 Curriculum standards and the newer Kurikulum Merdeka emphasize high-level thinking skills in learning outcomes, aligning with the main objective of developing 21st-century learning competencies.

Basic literacy that must be improved to face the challenges of the 21st century, one of which is numeracy literacy. Numeracy literacy is the ability to collaborate effectively with mathematical knowledge and understanding to face the challenges of everyday life by (1) using various symbols and numbers related to mathematics to solve problems in the context of everyday life, (2) analyzing information displayed in various representations (3) using the interpretation of the results of the analysis to predict and make decisions (Anjasari & Rahaju, 2025). Numeracy skills are used to help solve everyday contextual problems.

So far, the implementation of tests in schools is still widely carried out using paper (paper-based test). Implementing tests using paper has several disadvantages such as the high cost of printing questions and the length of time to correct and find out student learning outcomes (Zakaria et al., 2017) . The use of technology is not a barrier for students at SMPN 6 Bengkulu City, because the school allows students to use Android during the learning process. Research conducted by Kartini dan Putran (2020) also stated that there is a significant influence in the use of Android-based learning media on student learning outcomes.

Technological advances encourage the creation of evaluation media in the form of tests using computers and androids. Assessment using technology makes it easier for teachers and students because it is more economical, objective and transparent (Farman et al., 2020). However, in its implementation, many teachers still have difficulty in conducting assessments that utilize the use of technology (Kamilati, 2018). Research conducted by Mahfud et al. (2019) stated that almost 92% of 100 teachers did not know what technology could be integrated into the learning process, especially in making assessment instruments in the form of tests.

One of the technologies that can be used in creating assessment instruments is the ispring suite. Ispring Suite is an application device that can be used in creating test questions with various features such as creating many quiz models, inserting audio, video, or YouTube online or offline (Farman et al., 2021). Ispring suite provides many test options and the appearance of the final score after the assessment (Cahyanti et al., 2019). In its use, ispring suite is easy to use and does not take a long time to create, besides that ispring suite can also be collaborated with Microsoft Power Point so that it makes it easier for teachers to apply evaluation tests continuously (Afifah et al., 2021).

Although iSpring Suite has been extensively researched as a learning and assessment medium (Suprapti, 2016; Cahyanti et al., 2019), its application specifically for developing numeracy literacy test instruments—which require contextual problem-solving and complex reasoning characteristics—

remains limited. Previous studies primarily focused on general quiz creation and learning media development, while this research addresses the specific demands of numeracy literacy assessment aligned with AKM standards and PISA frameworks. The distinction is critical: numeracy literacy testing requires instruments that assess mathematical formulation, application, and interpretation in authentic contexts—requirements that differ substantially from general knowledge assessment.

Research conducted by Suprapti (2016) stated that the creation of learning devices using ispring suite meets the criteria for good learning devices. Likewise, research conducted by Cahyanti et al. (2019) stated that ispring suite can be used as a learning evaluation tool by creating test instruments using the features available in ispring suite. iSpring Suite offers distinctive advantages for numeracy literacy assessment compared to other digital platforms: (1) integration of multimedia elements (images, videos, diagrams) to present authentic contextual problems essential for numeracy assessment; (2) automatic scoring and immediate feedback mechanisms critical for diagnostic purposes; (3) randomization and security features ensuring test integrity; (4) cross-platform compatibility enabling seamless Android deployment. These features specifically support the three core processes of numeracy literacy: mathematical formulation (problem representation), application (solution finding), and interpretation of results in real-world contexts.

Technology use is not a barrier for students at SMP Negeri 6 Bengkulu City, as the school supports student access to Android devices during learning processes. Research by Kartini & Putran (2020) documented significant positive influence of Android-based learning media on student achievement. Given that smartphone penetration in Indonesia reaches 68% compared to computer ownership of 20%, Android-based assessment represents a practical and accessible solution for Indonesian educational contexts.

Based on the background analysis above, this research aims to develop an Android-based numeracy literacy test instrument using iSpring Suite that demonstrates validity, reliability, practicality, and strong item quality. The expected contribution includes providing teachers with a replicable model for developing digital numeracy literacy assessment instruments and advancing the national agenda of improving numeracy literacy through appropriate evaluation tools.

METHOD

Research Design

This study employs a Research and Development (R&D) approach within the ADDIE framework to develop an Android-based numeracy literacy test instrument using iSpring Suite. The purpose of developing this instrument is to create a valid, reliable, practical test instrument with good item quality for Grade VIII students at SMP Negeri 6 Bengkulu City. The ADDIE development model consists of five stages: Analysis, Design, Development, Implementation, and Evaluation (Cahyadi, 2019).

Population and Research Sample

The population in this study comprised all Grade VIII students at SMP Negeri 6 Bengkulu City. Purposive sampling was employed to select Class VIII B (n=30 students) based on the following criteria: (1) the class had completed the number patterns material in the curriculum, (2) all students had access to Android devices and demonstrated comfort using them, (3) the mathematics teacher confirmed students' readiness to participate in digital assessment, and (4) the class represented diverse academic abilities suitable for comprehensive item analysis. This sampling approach ensured that the selected class possessed characteristics necessary for valid and comprehensive instrument evaluation.

Data collection technique

This study employs multiple data collection techniques aligned with each phase of the ADDIE model:

Tabel 1
Multiple Data Collection Techniques Aligned With Each Phase of the ADDIE Model

Technique	Function	Data Collected	ADDIE Phase
Interviews	Identify learning needs, obstacles, and assessment practices	Qualitative information about existing assessment conditions and teacher perspectives	Analysis
Expert Validation Sheets	Assess content, construct, language validity, and media quality using 4-point Likert scale	Quantitative validity scores (1=Poor, 2=Fair, 3=Good, 4=Excellent); qualitative feedback for improvement	Development
Field Test (Student Assessment)	Obtain empirical data for item quality analysis	Student test responses and answers for validity, reliability, difficulty, and discrimination analysis	Implementation
Practicality Questionnaires	Evaluate ease of use, efficiency, and user satisfaction through 4-point Likert scale	Quantitative practicality scores from students and teachers	Implementation/Evaluation
Documentation	Record development process and product evolution	Photographs, screenshots, field notes, product iterations	All phases

Data Analysis Techniques

Data analysis is a research process that is carried out after all research data has been obtained completely. Data analysis needs to be carried out accurately so that conclusions from the research can be obtained clearly (Ali, 2016). Researchers in this study used several analysis techniques as follows:

Expert Validation

Expert validation is done qualitatively where the assessment will be done by interviewing experts. This expert validation is divided into two parts, namely validation of question items and validation of android application media.

Expert validation was conducted **quantitatively** using a 4-point Likert scale (1=Poor, 2=Fair, 3=Good, 4=Excellent) to assess: (1) Content Validity—whether test items accurately measure the defined indicators; (2) Construct Validity—whether item design follows established theoretical frameworks; (3) Language Validity—whether items use clear, age-appropriate Indonesian; (4) Media Validity—whether the Android interface is user-friendly and professionally designed.

Three expert validators (two mathematics education lecturers from University of Bengkulu and one ICT specialist from Dehasen University, Bengkulu) independently rated each aspect of the instrument. The validity coefficient was calculated using **Aiken's V formula**, with a minimum threshold of 0.75 for acceptance:

$$V = \frac{\sum s}{n(c-1)}$$

Where: s = rating minus minimum rating; n = number of raters; c = highest rating

Qualitative feedback through open-ended comments was also collected to guide product revision. This combination of quantitative and qualitative assessment ensures both statistical validity and practical improvement guidance.

Question Items
Validation of test items was carried out using interview techniques with expert validators. This interview was conducted to determine the content, construction, and language of the developed test instrument product. The validation results from this validator were used as a reference in revising the developed test items so that later the developed test items were valid for use. The validator's assessment was in the form of product improvement notes that had to be revised by the author until there were no more improvement notes from the validator and the product was said to be appropriate and could be used in the learning process. The following is a table regarding the validator's assessment of the developed questions.

Android Apps
Validation of android applications is done by validating the media with expert validators. Validation of this application is done to determine the practicality of the test instrument application that has been developed whether it is valid for use in the learning process or not. If the product is said to be valid, then this product can be used easily in the learning process. Conversely, if this product is not valid, the product must be revised and repaired until it is valid.

Empirical Test
Empirical testing is a data testing technique that is carried out using empirical data from trials in the form of quantitative data (Wulan & Rusdiana, 2015). The empirical test stages of the application of this test instrument include item testing and product practicality testing.

Question Items
Qualitative item analysis is the process of reviewing question items based on empirical data obtained from students' test answers in order to improve the quality of the questions. Aspects that need to be considered in this stage are that each question item is reviewed based on: level of difficulty, discriminating power, distractor function, reliability, and validity of the question (Fitrianawati, 2015). In this study, the aspects of the test items that were studied were validity, reliability, difficulty index, discrimination power, distractors, and students' conceptual understanding of the material on sequences and series.

Validity
The validity of the test items is carried out to test whether or not the questions tested are accurate in obtaining data. A test is said to be valid if the results are in accordance with the established criteria (Arikunto, 2018). The validity of this item is measured using the biserial correlation technique. The formula used is:

$$\gamma_{pbi} = \frac{M_p - M_1}{S_t}\sqrt{\frac{p}{q}}$$

Source :(Arikunto, 2018)

Items with $\gamma_{pbi} \geq 0.40$ are classified as valid.

Reliability
The reliability of the test instrument is related to the consistency of the test results. A good test instrument is an instrument that is consistent in providing data that is in accordance with reality (Wulan & Rusdiana, 2015). In a test, the test instrument used is said to be reliable if the measurement results obtained are close to the actual condition of the test participants (Retnawati, 2016). Reliability measurement is done using statistical calculations called the Reliability coefficient. To measure the Reliability coefficient, the following formula is used:

$$r_{11} = \left(\frac{n}{n-1}\right)\left(\frac{S^2 - \sum pq}{S^2}\right)$$

Source :(Arikunto, 2018)

Reliability coefficient $r_{11} > 0.70$ indicates high consistency.

Difficulty Index

The difficulty index is a measurement of the degree of difficulty of a question. A question is said to be good if the question has a balanced level of difficulty (Suryadi, 2020). The balance of these questions can be seen from the existence of easy, medium, and difficult questions proportionally in the questions being tested. A balanced comparison of questions is made with the comparison rule.$3-4-3$or for the number of easy, medium and difficult questions$3-5-2$ (Matondang, 2021)

The difficulty index of the questions is expressed in the form of an index (p)the size of which ranges from$0{,}00-1{,}00$ (Wulan & Rusdiana, 2015). Where the lower the difficulty index of the question, the more difficult the question is.Difficulty level analysis is carried out to measure the degree of difficulty of the questions developed.The formula used to calculate the question difficulty index is as follows:

$$P = \frac{B}{JS}$$

Source :(Arikunto, 2018)

Information :

P : Item difficulty index
B : Many students answered the questions correctly
JS : Total number of students who took the test

To determine the difficulty index of a question, it is grouped using the following criteria:

Table 2
Question Item Difficulty Index Indicators

Difficulty Index	Mark P
Difficult	$0{,}00 \leq P < 0{,}30$
Currently	$0{,}30 \leq P < 0{,}70$
Easy	$0{,}70 \leq P \leq 1{,}00$

Source :(Arikunto, 2018)

Based on the criteria above, an item is considered good if it has a difficulty index $0{,}30 \leq P < 0{,}70$with moderate criteria and has a comparative difficulty index, namelycomparison rules $3-4-3$or for the number of easy, medium and difficult questions, questions with easy and difficult criteria should not not be used but must be adapted to their use $3-5-2$ (Arikunto, 2018; Matondang, 2021).

Distinguishing Power

Distinguishing power is the ability of test items to differentiate between students who have mastered and those who have not mastered the material being tested (Fitrianawati, 2015). The ability of students is detected through the correlation coefficient or also called the discrimination index (D) of the question's distinguishing power. The higher the question's discrimination index, the more the question is able to distinguish students' abilities in the material being taught (Suryadi, 2020). In addition, the higher the discriminating power of the question, the better the question is at differentiating students' abilities (Wulan & Rusdiana, 2015).

The discrimination index of the differentiating power of questions is grouped based on the range of values.$0{,}00-1{,}00$. A low value range indicates that the question cannot be used to differentiate students' abilities, and vice versa, the higher the index obtained, the better the question.To calculate the discrimination index, the following formula is used:

$$D = \frac{B_A}{J_A} - \frac{B_B}{J_B} = P_A - P_B$$

Source :(Arikunto, 2018)

Classification: Poor (D < 0.20), Fair (0.20 ≤ D < 0.40), Good (D ≥ 0.40)

The Trickster

Distractors are possible answers that are incorrect in multiple choice objective tests (Fitrianawati, 2015). A distractor is said to function well if the distractor is chosen evenly across the question items. However, on the other hand, if the distractor is not chosen evenly, the distractor does not function well and needs to be replaced to improve the questions in the future (Suryadi, 2020).

Each distractor is said to function if it has a great attraction for students who do not understand the material to choose the answer option. Each distractor can be said to function if it is chosen at least 2% from all test participants (Matondang, 2021). Whether or not a distractor works can be measured using the following formula:

$$\text{IP} = \frac{\text{P}}{\left(\frac{N-B}{n-1}\right)} \times 100\%$$

Source :(Rukajat, 2018)

Test Instrument Android Application

Android applications that have been developed and tested by students will then be tested for the practicality of the Android application. The practicality of a product is the ease in terms of preparation, use, processing, interpretation, and administration (Rukajat, 2018). A product is said to be practical if the product is likely to be used in the learning process. A product is said to be practical if it meets the criteria of practicality, namely: (1) production does not take a long time and is easy, (2) does not require a lot of money, (3) assessment and processing of values is easy to do, and (4) implementation does not take a long time (Febriana, 2019).

The practicality testing of the android application was carried out using a Likert scale. The Likert scale is a scale used in measuring something, generally attitudes, opinions and perceptions (Maryuliana et al., 2016). The Likert scale has$4-5$question items that will form a value about the representation of individual traits such as attitudes and behavior. In assessing this scale, the average number of questions used will be taken. The table 2 is a table of Likert scale scores:

Table 3
Likert Scale

Score	Answer Options
1	Strongly disagree
2	Don't agree
3	Agree
4	Strongly agree

Source :(Maryuliana et al., 2016)

The stages carried out in the practicality analysis in this study are as follows:

1. Give a score for each Likert scale assessment item based on the Likert scale scoring guidelines, namely: (1) strongly disagree, (2) disagree, (3) agree, and (4) strongly agree (Maryuliana et al., 2016).
2. Calculate the total score given by students on the questionnaire using the following formula:

$$N = \frac{k}{N_k} \times 100\%$$

Source :(Tamba & Manurung, 2022)

Information :

N : Percentage score
k : Total score obtained
N_k : Maximum score

3. Match the results of student assessment scores with the practicality criteria based on the Table 4, the criteria at least practical.
4. Determining the level of practicality of the developed test instrument based on the assessment scores obtained.
5. Create a table of product practicality test results. The table of product practicality assessment results is presented.

Table 4
Product Practicality Criteria

Ability Criteria	Score Interval
Very Practical	$81{,}25\% < N \leq 100\%$
Practical	$62{,}5\% < N \leq 81{,}25\%$
Not Practical	$43{,}75\% < N \leq 62{,}5\%$
Very Impractical	$25\% < N \leq 43{,}75\%$

Source :(Tamba & Manurung, 2022)

RESULTS AND DISCUSSION

Results of Analysis Phase

The analysis stage is the first stage carried out in research with the ADDIE development model. This stage is carried out to identify learning problems and provide solutions to existing problems. The analysis stage is carried out by analyzing the needs of junior high school mathematics learning and curriculum analysis compared to learning outcomes. The results of the analysis stage will be explained as follows.

Analysis of Mathematics Learning Needs for SMP

The analysis of junior high school mathematics learning needs was conducted to determine the learning obstacles of students and the need for the provision of mathematics learning media. This activity was carried out by conducting interviews with mathematics teachers of SMP N 6 Bengkulu City. The results of the interview were that the questions tested on students were still routine questions and were not able to measure students' numeracy literacy abilities, so that many students had difficulty in solving numeracy literacy questions. This difficulty had an impact on the results of the numeracy literacy test which were quite low for Class VIII students of SMP N 6 Bengkulu City in the 2023/2024 academic year. Other problems found were that students had difficulty in identifying questions, changing questions into mathematical representations, reusing concepts and applying concepts to the problems given. In addition, the results of the AKM test of students of SMP Negeri 6 in literacy and numeracy were still very low. These low AKM test results indicate that improvements need to be made that can increase the numeracy literacy of these students.

Furthermore, questions related to formative tests, in developing and making test instruments, teachers often only take questions from books. So the problem is that the test only evaluates aspects of cognitive mastery of certain KD and cannot measure the extent of students' numeracy literacy abilities. Increasing numeracy literacy can be done by providing questions with the numeracy literacy type. This question is given based on suggestions given by the subject teacher. Providing questions about numeracy literacy is based on research by Warniasih & Susanti (2024) which also states that providing practice questions with the numeracy literacy type is continuously able to increase the numeracy literacy of the students. Thus, it is known the importance of giving tests that are in accordance with the numeracy literacy test so that they can increase students' numeracy literacy.

Test assessments are often conducted using paper (paper-based test). Assessments using paper have several disadvantages such as requiring large printing costs for test implementation, requiring a long time in the correction process, and being slow to determine students' abilities based on the learning outcomes they have worked on (Zakaria et al., 2017). This weakness causes the need for a long time for teachers to find out diagnostic information about students that will be used in designing learning. Therefore, other assessment media are needed that are more effective for test assessment. One of the media that can be used in implementing the test is the android application media.

Android application media has advantages over paper use, namely affordability, ease of use, rapid media development, and ease in the process of correcting the questions being tested. This media is affordable because teachers no longer spend money to prepare test sheets. Easy to use because students are used to using Smartphones in their daily routines. The media is developed quickly because teachers can immediately create media by adding questions to the ispring suite software, as well as ease in the correction process because the test results are immediately listed after students complete the test given.

The advantages of using this media make it easier for teachers to find out diagnostic information about students quickly after they have done the test.

Table 5
Interview findings revealed three critical gaps

Aspect	Existing Condition (Paper-Based Assessment)	Proposed Solution (Android iSpring)
Question Type	78% routine procedural items measuring basic calculation	Contextual numeracy literacy items measuring formulation, application, interpretation
Feedback Timeliness	3-5 days required for manual correction and feedback	Immediate automatic scoring and instant feedback
Item Analysis Process	Manual, time-consuming, limited diagnostic value	Automated via integrated SPSS functionality
Cost Efficiency	High printing and paper costs (~Rp 500,000 per assessment cycle)	Low digital distribution cost (~Rp 5,000 per student)
Accessibility	Limited to classroom with printed materials	Anywhere access via student personal Android devices
HOTS Measurement	Limited capability to measure higher-order thinking	Structured to assess reasoning-level competencies

Teacher interviews at SMP N 6 revealed that: (1) current test questions evaluated only cognitive mastery of specific competencies (KD), failing to measure students' broader numeracy literacy abilities; (2) students demonstrated difficulty in identifying problem requirements, transforming questions into mathematical representations, and applying concepts to novel contexts; (3) AKM (Asesmen Kompetensi Minimum/National Assessment) scores in numeracy literacy were substantially below national averages (specific scores pending availability).

The advantages of using this android application media have not been utilized optimally by teachers. In its implementation, tests using android applications are still difficult to develop for teachers. In fact, there are still many teachers who do not know that android applications can be used as a test implementation media.(Mahfud et al., 2019). In addition, most teachers still have difficulty in operating media that involves the use of technology.(Kamilati, 2018). The results of the interview at SMP N 6 Kota Bengkulu also showed that there has not been much development of assessment media based on android applications. Therefore, android application media was developed to add new knowledge for teachers in developing tests using android applications.

One of the media that can be used in developing assessment media using this android application is the ispring suite software. This software was chosen in line with research (Cahyanti et al., 2019)which states that ispring suite can be used as a medium for making good test instruments using the available features. The available features are in the form of making many quiz models, inserting audio, video, or YouTube links (Farman et al., 2020). And ispring suite can be collaborated with Microsoft Powerpoint which makes it easier to apply tests continuously.(Afifah et al., 2021).

Thus, the results of the learning needs analysis stage in the form of a numeracy literacy test instrument are needed. In addition, the test instrument is developed in the form of an android application by utilizing the ispring suite software as a test instrument development medium.

Curriculum Analysis

Curriculum analysis revealed that SMP N 6 Bengkulu City implements the Kurikulum Merdeka (Independent Curriculum), where numeracy literacy is positioned as a foundational competency assessed through Asesmen Nasional (National Assessment). For Grade VIII students (Phase D in Kurikulum Merdeka structure), the relevant Capaian Pembelajaran (CP—Learning Achievement) is: *"Students can identify patterns in number sequences and solve problems involving these patterns in various real-world contexts."* The developed instrument aligns with: - Reasoning Level numeracy literacy (demanding higher cognitive complexity) - Number/Quantity Content domain - Three core competency processes: 1. Mathematical formulation (representing real-world problems

mathematically) 2. Mathematical application (solving the mathematical representation) 3. Result interpretation (validating and applying solutions to original context). This alignment with both Kurikulum Merdeka and PISA/AKM frameworks ensures the instrument measures competencies emphasized in current Indonesian national assessments.

Design Phase Results

The design stage is the second stage carried out in research with the ADDIE development model. The design stage consists of several stages, namely the design of question items, the design of the android application, and the design of testing question items and android applications. The results of the design stage will be explained as follows.

Designing Test Items

The item design stage is carried out by collecting data on the questions to be developed. The outline of the developed test instrument is presented in Table 4 below.

Table 6

Test Instrument Grid

Item #	Learning Achievement (CP)	Content	Context	Cognitive Process	Assessment Focus
1	Making generalizations from patterns in number sequences	Numbers (Quantity)	Social (school context)	Reasoning	Identify pattern rules from visual/numerical data
2	Solving problems involving number sequence patterns	Numbers (Quantity)	Social (community context)	Reasoning	Apply pattern identification to word problems
3	Solving problems with number sequences and series	Numbers (Quantity)	Social (economic context)	Reasoning	Formulate and solve complex pattern problems
4	Solving problems with number sequences and series	Numbers (Quantity)	Social (organizational context)	Reasoning	Interpret and validate solutions in context

The steps for designing the draft questions consist of four steps. The first step is to determine the type of test, namely to diagnose students' numeracy literacy skills in the number pattern material. The second step is to determine the competency standards and test indicators as in Table 5.1. The third step is to determine the type of test, namely the multiple choice type with four answer options. The fourth step is to create a test instrument grid which is presented in Table 5.1.

Thus, 4 multiple-choice questions were obtained with 4 answer options at the reasoning level. The test instrument was in the form of multiple choices with a score of one for the correct answer and a score of zero for the wrong answer. The results of designing these questions were then used to design an android application.

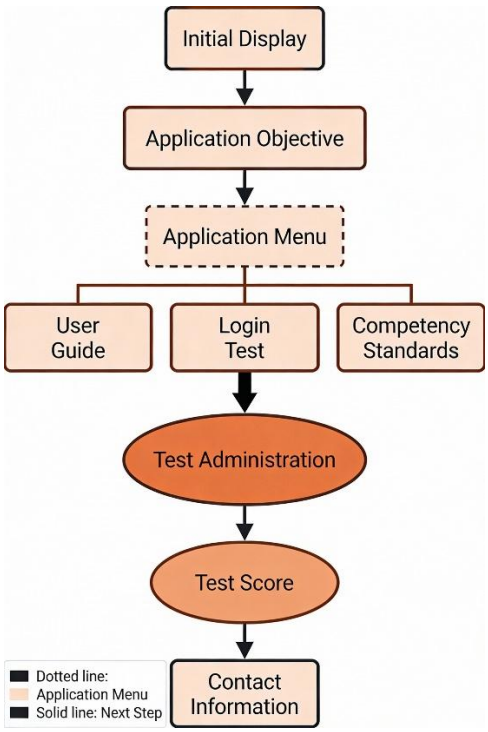
Android Application Design

The design of the android application is carried out after carrying out the analysis stage and aims to facilitate the process of developing an android application for a mathematical concept understanding ability test instrument. This design stage is carried out by creating a design for an android application for a numeracy literacy test instrument.

The android application design steps consist of three steps. The first step determines the flow of the media design developed. The media design flow is presented in Figure 1. The second step is

choosing the media title where the android application of electronic instruments based on the ARCS design model and android ispring suite is called the Numeracy Literacy Test Instrument application. The third step is determining the media layout. Determining the layout of the android application is used to design an attractive android application that is in accordance with the ARCS aspects so that it can increase students' learning motivation. There are three steps in determining the layout in the first android application, preparing the application that will be used in developing the android application, second collecting the media components used and finally determining the type of font and size that fits the android screen.

Figure 1
Android Application Usage Flow



In this way, the results of the Android application design were obtained which were used to design an Android application for test instruments that was attractive and in accordance with ARCS aspects. The design incorporated **ARCS (Attention, Relevance, Confidence, Satisfaction)** motivational design principles: - **Attention:** Colorful interface, animated transitions, clear typography - **Relevance:** Real-world contextual problems aligned with students' experiences - **Confidence:** Progressive difficulty, clear instructions, immediate feedback - **Satisfaction:** Achievement badges, score visualization, completion summary

Design of Test Items and Android Applications

Testing of test instruments is carried out logically and empirically to determine whether the questions developed are valid or not. Logical validation of test items is carried out to validators using construct content validity sheets, and language. Test items that have been logically tested are then empirically tested to students. After the test items are declared valid, the test items are then used in making android applications. Android applications are validated logically by validators using media validation sheets and empirically to determine the practicality of the application.

The preparation of the test instrument and application validation sheets is arranged based on aspects that are adjusted to the objectives of each validity. In the content validation sheet instrument, it consists of nine aspects, construction validity consists of six aspects, language validity consists of six aspects, and readability sheet consists of five aspects for each question item. While in the android application instrument, the media validation sheet consists of four aspects of the ARCS model and the

practicality sheet consists of four aspects of the use questionnaire. The list of validators in this study is presented in Table 7.

Table 7
List of Research Validators

Validity	Validator Jobs
Material, Construction, and Language	Lecturer of Mathematics Education, University of Bengkulu
	Teacher of SMP N 6 Bengkulu City
Media	Lecturer of Computer Education, Dehasen University, Bengkulu

The results of the test design stage are in the form of validation procedures for question items and Android applications, validation sheets, and validators involved in the research.

Development Phase Results
The development stage is the third stage carried out in research with the ADDIE development model. The development stage consists of several stages arranged based on the results of the analysis and design stages. The results of the development stage are explained in the section below.

Question Item Development
The data at the question item design stage is then arranged into draft question items that match the grid that has been prepared at the design stage. The results of the development stage are in the form of draft questions consisting of 4 multiple choice questions with 4 answer options. Four draft questions were developed and submitted for expert validation.

Tabel 8
Quantitative Expert Validation Results

Validation Aspect	Validator 1	Validator 2	Validator 3	Average Score	Aiken's V	Classification
Content Validity	3.89	3.78	4.00	3.89	0.96	Very Valid
Construct Validity	3.67	3.50	3.83	3.67	0.89	Very Valid
Language Clarity	3.75	3.67	3.92	3.78	0.93	Very Valid
Media Quality	3.83	4.00	3.92	3.92	0.97	Very Valid
Overall Average	3.79	3.74	3.92	3.82	0.94	Very Valid

Validation of Question Items
At this stage, the draft questions are validated by the validators. The validation carried out includes validation of the content, construct, and language of the question items. The validation results obtained that validator 1 had stated that the product was valid in terms of content, construct, and language because the product was in accordance with the validation sheet and no improvements were needed. Meanwhile, according to validator 2, the product had not been declared valid because there were still notes of improvements on the construct validity sheet for the aspect of using actual images/illustrations and clarifying the problem. Validator 1's improvement notes are shown in Table 9.

Table 9
Validator Revision Notes 1

Validity	Revised aspects	Validation Notes
Construct	Use of actual images/illustrations and clarify issues	In question number 4, the picture/illustration is clarified with picture instructions.

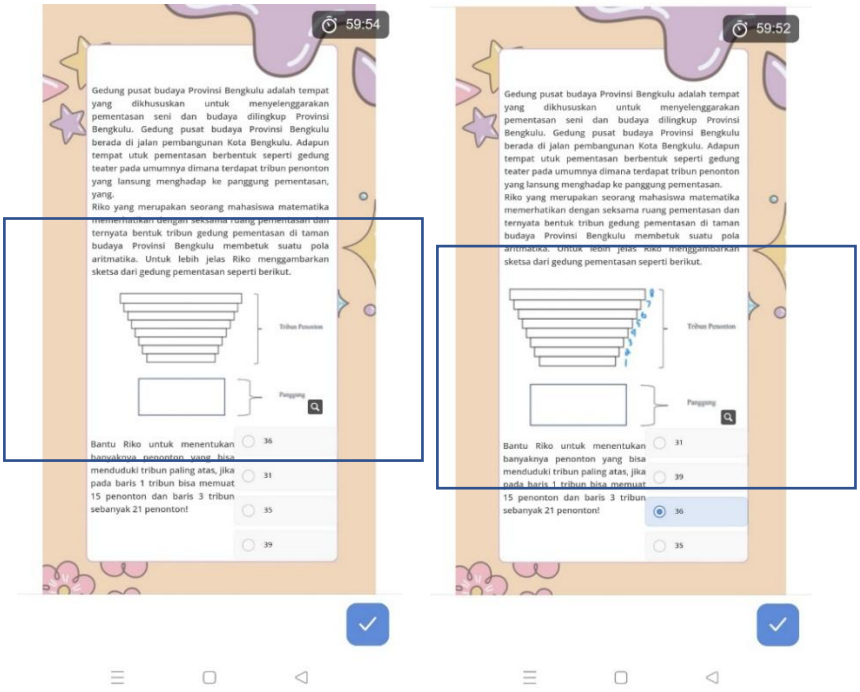
The revision notes were then used to re-arrange the questions that were more appropriate according to the assessment aspects of the material and construction validation. After the revision, the questions were then re-tested in the second validation by validator 1. In the second validation stage, the results of the material and construction validation by validator 1 were declared valid and met the established standards. Thus, the product was declared valid and ready to be used in the learning process.

Revision of Question Items

At this stage, the validation notes by the validator are used as material for revising the test items that can be used in diagnosing students' numeracy literacy skills. This stage is carried out based on the results of the validation notes by the validator in Table 9.

Figure 2

Revision Process



Before Revision After Revision

Revision of the image by providing a description of the sequence level. a related to questions that do not match the indicator restate the concept. Validator 1 said that in question number 4 the Image/Illustration is clarified with image instructions. This question item was then revised according to the validator's suggestion, namely providing a description of the sequence level in the image.

Field Test 1

At this stage, the test items that were declared valid by the validator were used for a trial in Class VIII B of SMP N 6 Kota Bengkulu. This field test was conducted on Thursday, August 15, 2024, involving 30 students. This field test aims to measure the validity, reliability, difficulty index, discriminating power, distractor function, and readability of each test item. Students will be given a test sheet that must be completed within 40 minutes, as well as a readability questionnaire sheet that is filled in after students complete the test. The results of the empirical test of the test items that have been carried out are explained in Table 10.

Validity test analysis is carried out to measure the ability of the test items to measure the test being developed.(Sukendra & Atmaja, 2020). Validity calculation using the biserial point correlation formula with the results of the calculation of the validity of the test items is shown in Table 10. In Table 10, it is obtainedThe validity value for all questions is in the range . This result shows that all questions are categorized as valid because they have met the minimum validity value.($\boldsymbol{\gamma_{pbis}}$)$0{,}440 \leq \gamma_{pbis} \leq 0{,}846 \geq 0{,}40$

Table 10
Results of the Validity Test of Question Items (Point-Biserial Correlation)

Number Question	Validitas (γ_{pbis})	V/T	Validity Criteria
1	0.482	Valid	Strong Enough
2	0.558	Valid	Strong Enough
3	0.526	Valid	Strong Enough
4	0.588	Valid	Strong

The validity of these questions is grouped into three criteria, namely very strong, strong, and quite strong. There is one question item with strong criteria, and the remaining 3 questions with quite strong criteria. Despite having variations in validity criteria, this instrument can still be relied on as a tool to measure understanding of mathematical concepts because it has met the valid criteria, namely$\boldsymbol{\gamma_{pbis}} \geq 0{,}40$.

Reliability test analysis is conducted to assess the consistency of test results. The results of the reliability analysis are obtained through the conformity between the expected score and the actual test results.(Setiawan, 2018). The test was conducted using the reliability coefficient formula 20 (KR-20) with a calculation result of 0.890. This reliability result shows that the instrument has been declared reliable because$r_{11} > 0{,}60$while the calculation results obtained data with very high reliable information. The calculation results show that this test can be relied on to measure the ability to understand mathematical concepts with high consistency.$0{,}890 > 0{,}60$

Analysis of the difficulty index and discriminating power of each question item was conducted to measure how difficult or easy a question is, as well as to measure how well a question is in distinguishing between participants with high and low abilities. Table 11 presents the results of the calculation of the difficulty index and discriminating power of the question items as follows.

Table 11
Results of the Difficulty Index (P) and Distinguishing Power (D) Test

Question Number	*P*	Criteria	*D*	Criteria
1	0.500	Currently	0.333	Medium question
2	0.633	Currently	0.467	Good question
3	0.233	Difficult	0.333	Medium question
4	0.200	Difficult	0.400	Medium Questions

The results of the analysis of the difficulty index and discriminating power in Table 8 obtained data on the difficulty index of 2 medium questions and two difficult questions. The discriminating power of question 2 is in the good question criteria, the rest have moderate discriminating power.

Distractor test analysis is conducted to evaluate the effectiveness of distractors in multiple-choice test questions based on the percentage level of students who choose the option. Distractors are said to function if they are chosen by 2% of the total students who take the test (Matondang, 2021). The results of the calculation of the percentage of students who chose the answer are presented in Table 12 below.

Table 12
Item Distractor Test Results

Item Question	Answer Options			
	A	B	C	D
1	10%	33%	50%	7%
2	37%	33%	7%	23%
3	7%	3%	7%	83%
4	63%	10%	7%	23%

Table 12 shows that the percentage range of students choosing each answer option is 3% – 83%It meansThe distractors on all questions functioned well with every answer option being selected.by students. These results show that the answer options for the questions have a certain level of clarity and appeal for students to choose these answer options.

Discussion

Validity of Electronic Test Instruments

The validity of the test instrument was assessed through expert validation and empirical testing of the test items and Android applications. Based on this validation, the instrument met the validity criteria in terms of material, construction, language, and media.

Validity of Question Items

The results of expert validation show that the test items meet the criteria for logical validity. The validity of the material is declared valid after revisions to several test items that do not match the indicators. This is relevant to Suryadi's research (2020), which states that material validity is achieved when the questions are able to measure the expected indicators.

The validity of the construct is also considered valid after revisions were made to the instructions for working on it and the appearance of the questions to make it more systematic. This opinion is in line with Sukendra & Atmaja (2020), who stated that the validity of the construct must be in accordance with the theoretical framework underlying the preparation of the instrument.

In addition, language validity is obtained through sentence revision to comply with good and correct Indonesian language rules. This adjustment supports the results of Setiawan's (2018) research, which emphasizes the importance of using easy-to-understand language to increase the effectiveness of the questions.

Validity of Android Applications

Validation of the Android application was conducted using the ARCS Design model approach (Keller, 1987), which includes aspects of attention, relevance, confidence, and satisfaction. The validation results show that the application attracts students' attention through interactive displays, is relevant to learning objectives, provides students with confidence through feedback, and satisfies users through automatic scoring features.

These results are consistent with the research of Susanti & Imbiri (2020), which found that ARCS Design-based applications are effective in increasing students' learning motivation. Thus, the validity of this application has met the relevant standards to support numeracy literacy learning.

Reliability and Feasibility of Question Items

The reliability of the test items was tested to ensure the consistency of the measurement results. Empirical tests on 5 test items showed a high level of reliability (reliability coefficient > 0.7). This finding is in line with Retnawati's research (2016), which states that reliable instruments are able to provide consistent measurement results in various conditions.

Difficulty index and the discriminatory power was also analyzed. The results showed that the distribution of questions consisted of 2 medium category questions, 2 easy category questions, and 1 difficult category question, with the discriminatory power of all questions being in the good category. These results support Arikunto's opinion (2018), which states that good questions must have an even distribution of difficulty to support students' abilities at various levels.

In the distractor aspect, all answer options were declared to function well because they were chosen by more than 5% of participants. This result is consistent with Matondang's (2021) research, which states that distractors are effective if they are able to attract participants' attention to choose them.

Practicality of Electronic Instruments

The practicality of the Android application was tested through a questionnaire involving users (students and teachers). The results showed that the application was easy to use, attractive, and shortened the test completion time. This opinion is in line with the findings of Harrati et al. (2016), which stated that the practicality of a product is influenced by ease of use and user satisfaction.

In addition, the use of Android-based applications as evaluation media is considered to support digital-based learning. Gan et al. (2015) also concluded that technology-based products provide learning variations and increase students' interest in solving problems.

The instruments developed in this study are in line with several previous studies. For example, Fitrianawati (2015) emphasized the importance of validity and reliability in developing questions for evaluating numeracy literacy skills. Meanwhile, the use of iSpring Suite for Android application development supports the results of Sulistyaningrum et al. (2024) research, which states that this software is effective for creating interactive and valid learning instruments.

CONCLUSION

This research successfully developed an Android-based numeracy literacy test instrument using iSpring Suite that met rigorous quality standards. Expert validation yielded an Aiken's V coefficient of 0.94 (very valid), while empirical testing demonstrated strong item validity ($\gamma_pbi \geq 0.482$), very high reliability ($r_{11} = 0.890$), and excellent practicality (85.2% user satisfaction). The instrument exhibited balanced difficulty distribution with adequate discriminating power (D = 0.333–0.467) and functional distractors, confirming sound psychometric properties .

This research addresses three critical gaps in Indonesian mathematics education: providing a replicable model for digital numeracy literacy assessment development, demonstrating alignment with Kurikulum Merdeka and AKM frameworks, and showing practical methods for mobile platform conversion to enhance accessibility in under-resourced schools. These contributions support teachers in creating context-appropriate assessment instruments while ensuring policy compliance and technological feasibility .

Important limitations include the small sample size (30 students from one school), the prototype scope (4 items for diagnostic rather than summative assessment), absence of longitudinal impact measurement, and limited regional generalizability beyond Bengkulu City. Future research should expand the item bank to 20-30 items, conduct multi-site validation across diverse Indonesian schools, investigate longitudinal learning impacts, explore adaptive testing features, develop teacher training materials, and compare effectiveness against traditional paper-based assessments .

This instrument advances digital evaluation media for numeracy literacy through both technical innovation (Android delivery) and pedagogical alignment (Kurikulum Merdeka and PISA frameworks). Despite scope limitations, the rigorous development process provides a validated prototype for Indonesian educators seeking to improve numeracy literacy assessment through effective technology integration.

REFERENCE

Afifah, W., Edy, S., & Huda, S. (2021). Pengembangan alat evaluasi pembelajaran matematika berbasis two tier multiple choice berlevel HOTS menggunakan ispring suite 9. *Postulat : Jurnal Inovasi Pendidikan Matematika*, *2*(2), 175–189. https://doi.org/10.30587/postulat.v2i2.3835

Ali, M. (2016). *Teknik Analisis Kualitatif*. Universitas Negeri Yogyakarta.

Anggalia, F., Bharata, H., & Rosidin, U. (2020). Developing PBL To Improve Mathematical Problem Solving And Self Efficacy. *International Journal of Trends in Mathematics Education Research*, *3*(1), 24–30. https://doi.org/10.33122/ijtmer.v3i1.129

Anjasari, T., & Rahaju, B. (2025). *Junior High School Students' Numeracy in Solving Number Content AKM Problems Based on Mathematical Ability*. *14*(1), 195–212. https://doi.org/10.26740/mathedunesa.v14n1.p195-212

Arikunto, S. (2018). *Dasar-Dasar Evaluasi Pendidikan* (3rd ed.). Bumi Aksara.

Budiman, I. A., Haryanti, Y. D., & Azzahrah, A. (2021). *Pentingnya Media Aplikasi Android Menggunakan Ispring Suite 9 Pada Pembelajaran Daring Terhadap Motivasi Belajar Siswa*. 144–150.

Cahyadi, R. A. H. (2019). Pengembangan Bahan Ajar Berbasis Addie Model. *Halaqa: Islamic Education Journal*, *3*(1), 35–42. https://doi.org/10.21070/HALAQA.V3I1.2124

Cahyanti, A. D., Farida, F., & Rakhmawati, R. (2019). Pengembangan alat evaluasi berupa tes online/offline matematika dengan ispring suite 8. *Indonesian Journal of Science and Mathematics Education*, *2*(3), 363–371. https://doi.org/10.24042/ijsme.v2i3.4362

Delima, N., Kurniasih, I., Tohari, Hutneriana, R., Amalia, F. N., & Arumanegara, E. (2022). *PISA dan AKM literasi matematika dan kompetensi numerasi* (Issue June).

Farman, F., Anjelina, S., Putri, Q. T., Mardiah, N. A., & Sari, K. (2021). Pengembangan instrumen penilaian pembelajaran matematika berbasis ispring suite. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *10*(4), 2040–2052. https://doi.org/10.24127/ajpm.v10i4.4288

Farman, Sarimuddin, & Chairuddin. (2020). Meningkatkan kemampuan guru dalam penyusunan soal evaluasi berbasis komputer menggunakan wondershare quiz creator. *SNPKM: Seminar Nasional Pengabdian Kepada Masyarakat*, *2*, 21–27.

Febriana, R. (2019). *Evaluasi pembelajaran*. Bumi Aksara.

Fitrianawati, M. (2015). Peran analisis butir soal guna meningkatkan kualitas butir soal, kompetensi guru dan hasil belajar peserta didik. *Prosiding Seminar Nasional Pendidikan PGSD UMS & HDPGSDI Wilayah Jawa*, *5*(3), 282–295.

Gan, B., Menkhoff, T., & Smith, R. (2015). Enhancing students' learning process through interactive digital media: New opportunities for collaborative learning. *Computers in Human Behavior*, *51*, 652–663. https://doi.org/10.1016/j.chb.2014.12.048

Harrati, N., Bouchrika, I., Tari, A., & Ladjailia, A. (2016). Exploring user satisfaction for e-learning systems via usage-based metrics and system usability scale analysis. *Computers in Human Behavior*, *61*, 463–471. https://doi.org/10.1016/j.chb.2016.03.051

Kamilati, N. (2018). Analisis komponen penilaian pada rencana pelaksanaan pembelajaran sebagai acuan pengembangan kurikulum diklat teknis sunstantif guru. *EDUKASI: Jurnal Penelitian Pendidikan Agama Dan Keagamaan*, *16*(1), 1–17. https://doi.org/10.32729/edukasi.v16i1.440

Kartini, K. S., & Putra, I. N. T. A. (2020). Pengaruh penggunaan media pembelajaran interaktif berbasis android terhadap hasil belajar siswa. *Jurnal Pendidikan Kimia Dan Ilmu Kimia*, *3*(2), 8–12. https://doi.org/10.33627/re.v3i2.417

Keller, J. M. (1987). Development and use of the ARCS model of instructional design. *Journal of Instructional Development*, *10*(3), 2–10. https://doi.org/10.1007/BF02905780

Lindawati, S. (2018). Literasi Matematika Dalam Proses Belajar Matematika Di Sekolah Menengah Atas. *Jurnal Prinsip Pendidikan Matematika*, *1*(1), 28–33. https://doi.org/10.33578/prinsip.v1i1.18

Mahfud, H., Purnama Adi, F., Widianto Atmojo, I. R., & Ardiansyah, R. (2019). Peningkatan kompetensi evaluasi pembelajaran berbasis teknologi pada guru SD di kota Surakarta. *Jurnal Pendidikan Dasar*, *7*(2), 1–5. https://doi.org/https://doi.org/10.20961/jpd.v7i2.35104

Maryuliana, Subroto, I. M. I., & Haviana, S. F. C. (2016). Sistem Informasi Angket Pengukuran Skala Kebutuhan Materi Pembelajaran Tambahan Sebagai Pendukung Pengambilan Keputusan Di Sekolah Menengah Atas Menggunakan Skala Likert. *Jurnal Transistor Elektro Dan Informatika (TRANSISTOR EI)*, *1*(2), 1–12.

Matondang, Z. (2021). *Evaluasi pembelajaran*. Program Pascasarjana Unimed.

OECD. (2023). PISA 2022 Results Factsheets Indonesia. *The Language of Science Education*, *1*, 1–9. https://oecdch.art/a40de1dbaf/C108.

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian* (1, Ed.). Parama Publishing.

Rukajat, A. (2018). *Teknik evaluasi pembelajaran*. Deepublish.

Setiawan, D. F. (2018). *Prosedur evaluasi dalam pembelajaran* (1st ed.). Deepublish.

Sukendra, I. K., & Atmaja, I. K. S. (2020). Instrumen penelitian. In *Deepublish*. Mahameru Press.

Sulistyaningrum, H., Nuraida, D., Wardhono, A., Andik, M., & Author, C. (2024). The Effectiveness of Use of Power Point-Ispring Interactive Learning Media on Student Creativity. *Journal of Humanities and Social Sciences Studies*, *6*(3), 83–90. https://doi.org/10.32996/jhsss

Suprapti, E. (2016). Pengembangan Perangkat Pembelajaran Matematika Model Kooperatif Tipe STAD dengan Media Powerpoint Ispring pada Materi Jajargenjang, Layang-Layang, dan Trapesium Di Kelas VII SMP. *MUST: Journal of Mathematics Education*, *1*(1), 57–68.

Suryadi, A. (2020). *Evaluasi Pembelajaran* (Jilid II). CV Jejak.

Suryapuspitarini, B. K., Wardono, & Kartono. (2018). Analisis Soal-Soal Matematika Tipe Higher Order Thinking Skill (HOTS) pada Kurikulum 2013 untuk Mendukung Kemampuan Literasi Siswa. *Prisma, Prosiding Seminar Nasional Matematika*, *1*, 876–884.

Susanti, L., & Imbiri, C. (2020). Implementasi Motivasi Model ARCS (Attention, Relevance, Confidence, and Satisfaction) dalam Pembelajaran Pendidikan Agama Kristen. *DUNAMIS: Jurnal Teologi Dan Pendidikan Kristiani*, *4*(2), 254–263. https://doi.org/10.30648/dun.v4i2.284

Tamba, J. S., & Manurung, N. (2022). Pengembangan media pembelajaran menggunakan articulate storyline 3 untuk meningkatkan kemampuan pemahaman konsep matematis siswa pada materi bangun ruang sisi datar di SMP Swasta Katolik Budi Murni 2 Medan. *Inspiratif : Jurnal Pendidikan Mtematika*, *8*(2), 52–66. https://doi.org/https://doi.org/10.24114/jpmi.v8i2.37196

Warniasih, K., & Susanti, E. (2024). Improving Math Learning Outcomes by Implementing Numeracy Literacy Drill Exercises. *International Conference on Aplied Social Sciences in Education*, *1*(1), 667–676. https://doi.org/10.31316/icasse.v1i1.6892

Wulan, E. R., & Rusdiana, H. A. (2015). Evaluasi pembelajaran dengan pendekatan kurikulum 2013. In *Pustaka Setia*. Pustaka Setia.

Zakaria, Hadiarti, D., & Fadhilah, R. (2017). Pengembangan instrumen evaluasi berbasis CBT dengan software ispring quizmaker pada materi kesetimbangan kimia. *Jurnal Pendidikan Matematika Dan Sains*, *1866*(2), 178–183. https://doi.org/10.21831/JPMS.V5I2.16709

Zulfayani, Z., Ariawan, R., Nufus, H., Nurdin, E., & Fitriani, D. (2024). Pengembangan instrumen kemampuan literasi numerasi berbasis HOTS. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *8*(1), 789–800. https://doi.org/10.31004/cendekia.v8i1.2678