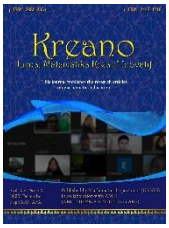
UNNES JOURNALS

Kreano, Jurnal Matematika Kreatif-Inovatif
Volume 16, number 1, 2025, pp. 44-61



Development of Realistic Mathematics Education-based Numeracy Test Instrument with Maritime Context for Junior High School Students in Banda Aceh, Indonesia

Bintang Zaura[1], Siti Fatimah[2,3], Suhartati[1], Mukhlis Hidayat[1], Nurul A'la[1], Fahlida Harnita[2]

[1]Department of Mathematics Education, Universitas Syiah Kuala, Aceh, Indonesia 23111
[2]PRP-PMRI (Center for Research and Development of Indonesian Realistic Mathematics Education)
[3]Doctoral Program in Elementary Education, Universitas Negeri Malang, Malang, Indonesia 65145

Correspondence should be addressed to Siti Fatimah:
siti.fatimah.2521039@students.um.ac.id

Abstract

Indonesia has changed its national student assessment approach, focusing on literacy-based assessment in 2021. Numeracy skills are a critical component of this framework, requiring valid, practical, and effective instruments for their evaluation. This study investigates the development and feasibility of numeracy test instruments contextualized within a maritime framework and grounded in Realistic Mathematics Education (RME). Using Tessmer's Research and Development (R&D) method, the study followed four stages: self-evaluation, expert review, one-to-one testing, small group testing, and field tests. A total of 25 seventh-grade students with varying mathematical abilities participated in the study. Data were analyzed descriptively using observations, tests, and document analysis. The findings revealed that the developed instruments are valid, practical, and effective. The maritime context successfully integrates real-world relevance into mathematics learning, fostering numeracy skills. These instruments provide educators with innovative tools to connect mathematics with environmental and regional issues, such as marine conservation, enhancing both numeracy and interdisciplinary understanding. This research underscores the importance of contextualizing numeracy assessments and offers a replicable framework for future applications. The study's findings imply that the developed instruments can serve as an educational model for other coastal regions in Indonesia, enhancing numeracy skills while promoting awareness of maritime environmental issues. In conclusion, the RME-based maritime numeracy instrument is ready for classroom use and provides a replicable model for coastal regions seeking to enhance numeracy while fostering marine-environment awareness.

Keywords: Numeracy Skills; RME; Instrument Development; Maritime Context.

Information of Article

Subject classification	*97C40 Assessment (large scale assessment, validity, reliability, etc.)*
Submitted	*23 August 2024*
Review Start	*24 August 2024*
Round 1 Finish	*15 May 2025*
Round 2 Finish	*17 June 2025*
Accepted	*23 June 2025*
Scheduled online	*30 June 2025*
Similarity Check	*6%*

Abstrak

Indonesia telah mengubah pendekatan penilaian siswa nasional, dengan fokus pada penilaian berbasis literasi pada tahun 2021. Kemampuan numerasi merupakan komponen penting dalam kerangka kerja ini, yang membutuhkan instrumen yang valid, praktis, dan efektif untuk evaluasinya. Penelitian ini menyelidiki pengembangan dan kelayakan instrumen tes numerasi yang dikontekstualisasikan dalam kerangka kerja kemaritiman dan didasarkan pada Pendidikan Matematika Realistik (RME). Dengan menggunakan metode Penelitian dan Pengembangan (R&D) dari Tessmer, penelitian ini mengikuti empat tahap: evaluasi diri, tinjauan ahli, uji coba satu lawan satu, uji coba kelompok kecil, dan uji coba lapangan. Sebanyak 25 siswa kelas tujuh dengan berbagai kemampuan matematika berpartisipasi dalam penelitian ini. Data dianalisis secara deskriptif dengan menggunakan observasi, tes, dan analisis dokumen. Hasil penelitian menunjukkan bahwa instrumen yang dikembangkan valid, praktis, dan efektif. Konteks maritim berhasil mengintegrasikan relevansi dunia nyata ke dalam pembelajaran matematika, sehingga menumbuhkan kemampuan berhitung. Instrumen-instrumen ini menyediakan alat inovatif bagi para pendidik untuk menghubungkan matematika dengan isu-isu lingkungan dan regional, seperti konservasi laut, sehingga meningkatkan kemampuan berhitung dan pemahaman interdisipliner. Penelitian ini menggarisbawahi pentingnya kontekstualisasi penilaian numerasi dan menawarkan kerangka kerja yang dapat ditiru untuk aplikasi di masa depan. Temuan penelitian ini menyiratkan bahwa instrumen yang dikembangkan dapat berfungsi sebagai model pendidikan untuk wilayah pesisir lainnya di Indonesia, meningkatkan kemampuan numerasi. Kesimpulannya, instrumen numerasi berbasis PMR-kemaritiman ini layak digunakan di kelas dan dapat direplikasi di wilayah pesisir lain untuk meningkatkan numerasi sekaligus kesadaran lingkungan laut.

INTRODUCTION

Indonesia's geopolitical identity is inextricably bound to the sea: more than 70 % of its territory is maritime, comprising over 17 000 islands, 108 000 km of coastline, and vast exclusive economic zones that rank among the largest in the world (Ebarvia, 2016). The nation enjoys stewardship over the Coral Triangle's richest reef systems, mangrove forests that sequester more blue carbon than any other country, and fisheries that sustain millions of coastal households. Policy documents such as the "Visi Maritim Nusantara 2045" predict that a fully realised blue economy could contribute at least a quarter of Indonesia's gross domestic product by mid-century, generating upwards of ten million new jobs in sustainable aquaculture, eco-tourism, and marine renewable energy. Yet these ambitions are threatened by persistent illegal, unreported, and unregulated (IUU) fishing, habitat degradation, and climate-induced hazards that jeopardise both biodiversity and coastal livelihoods (Chapsos & Hamilton, 2019). Tackling such multifaceted challenges demands a citizenry equipped not only with environmental ethics but also with the numerical competencies required to interpret data, model risk, and make evidence-based decisions about marine resources.

Contemporary educational discourse therefore frames numeracy as far more than elementary computation; it is the capacity to *critically interpret, communicate, and act upon quantitative information in diverse, authentic contexts*. International assessments such as the Programme for International Student Assessment (PISA) define mathematical literacy as the foundation for full participation in an information-rich society. Closer to home, the Indonesian government's 2021 overhaul of national examinations into the Asesmen Kompetensi Minimum (AKM) elevates numeracy—alongside reading literacy—to a core metric of school quality and district accountability (Simarmata & Mayuni, 2023). The reform aspires to spur instructional shifts from rote calculation to problem solving, reasoning, and data handling—skills that align directly with the cognitive demands of a modern

maritime economy. However, realising this vision presupposes the availability of valid, reliable, and contextually relevant assessment instruments that can diagnose students' competencies and guide targeted remediation. Existing item banks remain dominated by abstract or urban-centric scenarios (e.g., sharing pizza or calculating taxi fares) that fail to resonate with the lived experiences of the roughly 40 % of Indonesians residing in coastal districts. The resulting disconnect risks underestimating actual ability, misinforming instructional planning, and undermining student motivation.

Realistic Mathematics Education (RME), first conceptualised in the Netherlands and adapted globally, offers a theoretically robust pathway for designing assessments that bridge school mathematics with everyday experience. RME rests on two pillars: *guiding reinvention*—students progressively formalise informal strategies—and the *reality principle*—meaningful contexts anchor the mathematics to be learned (Gravemeijer, 1998; Treffers & Goffree, 1985). Empirical syntheses show that RME-aligned tasks foster deeper conceptual understanding, perseverance, and transfer, particularly when they mirror learners' socio-cultural realities (I. D. Lestari et al., 2024; Uyen et al., 2023). Coastal Indonesia offers a wealth of naturally mathematical phenomena—tidal oscillations, boat-navigation angles, fish-stock quotas, coral-reef growth rates—that can instantiate RME's principles while simultaneously cultivating ecological stewardship (Palinussa et al., 2025). Yet a review of Indonesian numeracy assessments published since 2015 reveals that fewer than 5 % incorporate maritime or ecological content, and fewer still undergo rigorous psychometric validation.

Globally, scholars argue that decontextualised testing regimes fall short of measuring twenty-first-century competencies. Sriraman and English (2010) contend that creativity, collaboration, and resilience emerge most authentically when learners confront mathematics embedded in real, socio-scientific narratives. Contextualised tasks leverage prior knowledge, reduce extraneous cognitive load, and honour cultural assets, thereby advancing equity for students historically marginalised by one-size-fits-all examinations. Indonesian research echoes this insight. Istiana et al. (2021) integrated maritime texts into literacy instruction and reported gains in comprehension and engagement, yet their accompanying mathematics tests remained conventional. Wijayati et al. (2021) used coastal-resource simulations to bolster entrepreneurial attitudes but again assessed numeracy with generic items. Alfredo (2023) reviewed inquiry worksheets on marine pollution but found no aligned test instruments, underscoring a systemic gap between innovative pedagogy and assessment practice. Without psychometrically sound tools that capture context-specific reasoning, teachers cannot verify learning outcomes, and policymakers cannot benchmark progress across Indonesia's diverse archipelagic regions.

Addressing this gap is both timely and urgent for Banda Aceh, a city whose socioeconomic history is deeply entangled with the sea—from pepper-trading sultanates to modern tuna export hubs. The 2004 Indian Ocean tsunami tragically revealed how deficiencies in numerical risk literacy—understanding probability curves, wave-height projections, and early-warning signals—can exacerbate disaster impacts. More recently, rising sea-surface temperatures and plastic debris threaten key fisheries

and tourism corridors, amplifying the need for numerate citizens who can participate in coastal-management dialogues. Embedding such real challenges inside assessment tasks offers dual benefits: it gauges students' quantitative proficiency while simultaneously cultivating attitudes and knowledge that underpin sustainable ocean governance. Yet these very challenges complicate design; tasks must balance realism with cognitive accessibility, integrate interdisciplinary knowledge without overwhelming novice learners, and yield scoring rubrics that capture nuanced reasoning rather than mere procedural fluency.

Equity considerations further heighten the stakes. Indonesia's coastal zones encompass urban centres with high digital connectivity and remote islands where instructional resources are scarce. Language diversity adds complexity; even within Aceh, terms like *"gelombang pasang"* (storm surge) and *"biodegradable waste"* may not appear in students' everyday lexicon. Tests must therefore be pilot-tested for linguistic clarity, visual accessibility, and differential item functioning (DIF) across gender, ability, and socioeconomic lines. Failure to do so risks reinforcing educational disparities and undermining the broader goals of the AKM reform.

Tessmer's formative Research-and-Development (R&D) model offers a rigorous yet flexible framework to navigate these design challenges. Its five iterative stages—self-evaluation, expert review, one-to-one cognitive labs, small-group trials, and field testing—promote continual refinement based on empirical evidence of content validity, cognitive demand, and learner uptake (Zulkardi et al., 2020). Indonesian applications of Tessmer have produced reliable probability modules (Armiati et al., 2022) and COVID-context geometry tasks (Zulkardi et al., 2020); each reports marked improvements in item discrimination and engagement after cycling through expert and student feedback. Nonetheless, none have foregrounded maritime numeracy, nor have they leveraged cross-disciplinary panels that include marine scientists to vet the ecological fidelity of scenarios. Moreover, prior studies seldom report think-aloud analyses or cognitive-load measures, both critical for ensuring that language or graphics do not unintentionally elevate item difficulty.

Against this backdrop, the present study pursued three interconnected objectives. First, we designed a suite of 16 numeracy-test items grounded explicitly in Banda Aceh's maritime realities: industrial waste effluent, plastic-waste accumulation, oil-spill dispersion, and destructive fishing practices. Each item was mapped to Indonesia's AKM numeracy descriptors encompassing number sense, measurement, geometry, and data analysis. Second, the items underwent iterative refinement through a cross-disciplinary expert panel—three mathematics-education scholars, two marine-science academics, and three practising junior-high teachers—followed by one-to-one and small-group trials with students representing high-, medium-, and low-ability bands. This process was designed to triangulate technical metrics (content validity indices, Aiken's V) with qualitative evidence (student think-aloud protocols, engagement ratings). Third, we implemented a field test with 25 Grade VII students to evaluate the instrument's psychometric properties—difficulty, discrimination, internal consistency—and its practicality for 3 × 40-minute classroom administration.

The decision to focus on seventh-grade learners is strategic. At this

juncture, students transition from concrete operational to early formal-operational thinking, making them receptive to contextual bridges between lived experience and abstract representation. Early diagnostic information can guide teachers in tailoring instruction before students encounter the high-stakes AKM numeracy examination in Grade VIII. Furthermore, adolescence is a critical window for shaping pro-environmental attitudes; exposure to authentic marine issues during this formative period can engender habits of mind that persist into adulthood. By aligning tasks with national competency descriptors while embedding local content, the study responds simultaneously to policy mandates and community realities, offering a replicable template for other coastal regions.

The theoretical contributions of this study extend RME research into the underexplored realm of *assessment design*. While countless studies demonstrate RME's power in instructional settings, fewer interrogate whether its core principles can coexist with the rigour demanded by psychometric standards. By demonstrating that authentic, context-rich items can achieve satisfactory reliability (KR-20 ≥ 0.80) and item-fit statistics, we challenge the assumption that realism must be sacrificed for measurement precision. Additionally, we advance the emergent concept of "blue numeracy," defined here as the intersection of quantitative reasoning and ocean literacy, echoing global initiatives such as UNESCO's Decade of Ocean Science for Sustainable Development (2021–2030), which calls for integrating ocean knowledge across disciplines, including mathematics.

Practically, the study produces a ready-to-deploy instrument that educators can use for formative assessment, summative benchmarking, or as anchor tasks within project-based learning. Because scenarios draw on spectacles students witness—beach clean-ups, tidal-height charts posted at fishing docks, news reports of oil spills—they promise higher engagement and potential transfer to everyday decision making. For district and provincial administrators, psychometrically vetted items provide evidence to allocate professional-development resources more strategically, identify curricular blind spots, and benchmark progress toward numeracy targets in the AKM.

The research also models a collaborative design process bridging disciplinary silos. Mathematics educators contributed expertise on cognitive progression and item formats; marine scientists vetted ecological realism and supplied up-to-date datasets; classroom teachers ensured linguistic appropriateness and classroom feasibility. This triangulation embodies the "co-design" ethos central to contemporary curriculum-reform movements, illustrating how practitioner-scholar partnerships can accelerate the production of culturally and ecologically grounded assessments.

In summary, integrating maritime realities into numeracy assessment aligns with Indonesia's dual imperatives of educational excellence and sustainable ocean stewardship. The literature highlights both the pedagogical promise and the practical paucity of such endeavours, positioning the present study to fill a pressing void. By grounding Realistic Mathematics Education within Banda Aceh's coastal context and subjecting the resulting instrument to rigorous formative evaluation, we aim to contribute to three domains: (1) empirical evidence for context-based assessment design, (2) policy discourse on

regionalised item banks that preserve national comparability, and (3) classroom practices that nurture mathematically literate, environmentally conscious citizens equipped to navigate the complexities of a maritime nation in the twenty-first century.

METHOD

Research Design

The present study adopted a design-based research (DBR) orientation embedded within Tessmer's formative Research-and-Development (R&D) cycle. DBR was chosen because it combines systematic inquiry with iterative refinement, allowing researchers to move back and forth between theoretical propositions and classroom realities while documenting how successive design decisions influence learning outcomes. In line with best practice in educational DBR (McKenney & Reeves, 2019), each phase produced design artefacts—item blueprints, scoring rubrics, cognitive-load checklists—that were critically appraised before the next phase commenced.

Grade VII students (12 – 13 years) constituted the focal population for three reasons. First, the Indonesian Minimum Competency Assessment (MCA) calendar schedules the first high-stakes numeracy test at Grade VIII, making Grade VII an optimal window for early diagnostic feedback. Second, cognitive-development literature positions early adolescence as a transitional stage from concrete operations to formal reasoning, so contextualised tasks introduced now can accelerate abstraction. Third, the local curriculum devotes a full semester to ratio, proportion, and data handling—constructs that align naturally with maritime phenomena such as tide charts and fish-catch tallies.

Participants

The study ultimately involved two mutually reinforcing participant groups:

1. Students (n = 25) from a coastal public junior-high school in Banda Aceh. A stratified purposive approach ensured representation of high- (top 30 %), medium- (middle 40 %), and low-ability (bottom 30 %) learners based on their most recent mathematics report-card grades. This stratification was critical for stress-testing item difficulty and discrimination indices across ability bands.
2. Expert validators (n = 8) comprising three mathematics-education lecturers, two marine-science lecturers, and three practising junior-high teachers (one of whom specialised in biology to vet ecological accuracy). Their interdisciplinary composition was intended to safeguard both psychometric rigour and thematic authenticity.

Procedure

1. Preliminary Stage

A needs analysis triangulated three data sources: (a) MCA blueprints for numeracy, (b) Aceh Province syllabus documents, and (c) semi-structured interviews with two mathematics supervisors to probe existing assessment gaps. A concept map was then drafted to align MCA descriptors—e.g., "interpret data in tables/graphs," "solve ratio problems"—with maritime activities familiar to local youth, such as estimating the volume of plastic waste collected during monthly beach clean-ups. Sixteen prototype items were generated, sequenced from lower to higher levels of cognitive demand using Bloom's revised taxonomy.

2. Expert Review and One-to-One Testing

An initial prototype of the numeracy test was developed based on the Minimum Competency Assessment (MCA) framework. Questions were designed to reflect maritime issues such as industrial waste, environmentally harmful fishing practices, and oil spills.

Expert review sessions involved eight validators who assessed the prototype for clarity, relevance, and alignment with RME principles. Feedback was incorporated into revisions, ensuring content accuracy and contextual relevance.

One-to-one testing was conducted with three students (high, medium, and low ability) to assess readability and comprehension. Observations and student feedback informed further refinements.

3. Small Group Testing

The revised prototype was tested with six students, representing diverse mathematical abilities. This stage evaluated the practicality and usability of the instrument, with particular focus on question clarity, problem-solving strategies, and student engagement.

4. Field Testing

The final instrument was administered to 25 students during a 3-session (3×40 minutes) assessment. This stage evaluated the instrument's effectiveness in measuring numeracy skills, categorized as very high, high, medium, low, or very low.

Data Collection Techniques

Three primary data collection methods were employed:

1. Observations: Monitored student responses and interactions during testing.
2. Tests: Collected quantitative data on student performance across various numeracy questions.
3. Documentation: Analyzed student responses, expert feedback, and revisions made at each stage.

Data Analysis

Data analysis was carried out by analysing the results of validation by experts, one-on-one, small group, and field trials and used to revise test instruments made by researchers. The data were analysed through a qualitative descriptive method to describe the results of each stage of development in this study. Data analysis of the trial results was then used to determine the potential effect of students' numeracy skills by providing numeracy instruments.

The data processing of this research is by calculating the percentage of respondents' answers with the aim of seeing the comparison of the size of the frequency of respondents' answers on each different item. The formula used:

The validity test was conducted by a validator. Giving an assessment based on a Likert scale. The alternative answers given are as follows: 5 = Very good, 4 = Good, 3 = Neutral, 2 = Less, and 1 = very less (Borich, 2016). After the data is collected, the average score is calculated through the formula:

$$\bar{x} = \frac{\sum x}{N}$$

where $\bar{x}$ = average value of the score; $\sum x$ = number of values; and N = number of indicators.

The average assessment is based on the table of criteria for the validity of the instrument developed, so that conclusions can be obtained about the quality of the question instrument. The

criteria table is compiled based on the highest (ideal) score = 5, lowest score (1), number of classes/classifications = 5, and interval distance 0.8.

Based on the calculations, the validity criteria table is obtained with an interval distance of 0.8 as presented in the Table 1.

Table 1. Criteria for Validity of the Instrument

Range of average score	Criteria
>4,20 s/d 5,00	Very Feasible
>3,40 s/d 4,20	Feasible
>2,60 s/d 3,40	Less Feasible
>1,80 s/d 2,60	Not Feasible
1,00 s/d 1,80	Highly Not Feasible

RESULT AND DISCUSSION

Results

Observations and Preliminary Findings

Initial classroom walk-throughs and informal interviews with two mathematics teachers revealed that maritime examples were rarely incorporated into daily instruction. During a typical 40-minute lesson devoted to ratio and proportion, teachers relied on textbook scenarios (e.g., mixing paint or sharing pizza) rather than drawing on local coastal life. Field notes showed that of 39 questions completed by students in regular classwork, only two referenced any real-world context, and neither related to maritime issues. This observation confirmed survey data from the Aceh Provincial Education Office indicating that fewer than 15 % of junior-high mathematics teachers had attended professional-development sessions on contextualised learning in the past three years.

A diagnostic pre-test (ten items unrelated to the study instrument) administered one week before the project began placed students into three achievement bands: high (scores ≥ 80; n = 8, 32 %), medium (60 ≤ scores < 80; n = 9, 36 %), and low (scores < 60; n = 8, 32 %). These proportions were later used as sampling quotas for the one-to-one and small-group phases. Notably, the pre-test contained a single maritime problem ("Estimate the daily catch of a fisherman who brings home 24 kg on Monday, 18 kg on Tuesday … "). Only 20 % of students answered it correctly, supporting the premise that authentic contexts, when unfamiliar in assessment format, pose difficulties even to high achievers.

Development of the Test Instrument

The 16 prototype items were organised into four thematic clusters—Industrial Waste, Plastic Pollution, Oil-Spill Dispersion, and Destructive Fishing Gear—each mapping onto at least two MCA content descriptors. Item 5, for example, required proportional reasoning to scale up the daily discharge of a seafood-processing plant from "per hour" to "per week," while Item 12 asked students to interpret a multi-line graph depicting annual trends in plastic debris collected by local NGOs. Table 2 (unchanged heading) lists representative revisions; here we add a brief explanation of design logic: icons and photographs accompanying each stem were deliberately chosen from Banda Aceh news outlets and government reports to heighten authenticity. Every item was further tagged by cognitive-process level (CPL) using Bloom's revised taxonomy; six items reached CPL 4 or above (analyse, evaluate, or create).

Formative Evaluation

1. Expert Review and Validation

Eight validators, comprising marine science and mathematics education experts, evaluated the initial prototype to

ensure its accuracy and effectiveness. Based on their feedback, several revisions were made to enhance its quality. These included ensuring the consistent use of symbols and scientific terms, simplifying the language to improve readability, and adjusting image resolutions for better clarity. Additionally, decimal numbers were replaced with whole numbers to simplify calculations, and questions addressing climate change and its impact on the marine environment were added. These modifications, as detailed in Table 2, aimed to create a more user-friendly and comprehensive resource.

Table 2. Revised Questions Test

Prior Design	Validated Version
The inconsistent use of percentage symbol. Part of student used word "percent" and other used symbol	It has been revised and suggested to use the symbol "%"
The use of foreign language shall be in italic.	The use of foreign language has been revised. For example, the question No.7: National Oceanic and Atmospheric Administration (NOAA) has been written as *National Oceanic and Atmospheric Administration (NOAA)*
The use of sentences that integrates the science lessons should be given an explanation. For example, question no. 1 used the word "excretion"	In question no. 1, the caption "excretion (disposal of metabolic waste) is added. In question no. 7, "oil spill" was explained to be oil spill.
Image resolution needs to be considered, especially if it contains information to answer questions.	The image has been clarified from the color and use of captions on the image.
The use of decimal numbers needs to be considered especially in questions that take time to solve. In fact, students have used a lot of time to read the literacy of the questions which are also quite long.	Decimal numbers have been converted into whole numbers, to make it easier for students to calculate.
The maritime context in each question content is very good to increase students' knowledge about marine and marine environmental pollution.	The researcher designed a questionnaire about students' awareness of marine environmental pollution.

First, the prototype displayed inconsistent notation: some items spelled out "percent," whereas others used the symbol "%," and temperature readings alternated between "C" and "°Celsius." Although seemingly minor, such irregularities can distract students and inflate working-memory load, particularly for those in the low-achievement band. The revised version therefore standardises all mathematical and scientific symbols to the most compact, internationally recognised forms—"%," "°C," "kg"—thereby freeing cognitive resources for reasoning rather than decoding.

Second, foreign acronyms and institutional names (e.g., NOAA, FAO) appeared without formatting cues, making them look like undefined variables. Following APA style, every non-Indonesian term is now italicised on first mention and accompanied by its full English expansion (e.g., *National Oceanic and Atmospheric Administration [NOAA]*). This convention not only aids pronunciation but also signals source credibility, an important facet of data-

literacy instruction.

Third, several stems embedded specialised science vocabulary—*excretion*, *bioaccumulate*, *noxious*—that triggered confusion during think-aloud interviews. Rather than eliminating these authentic terms and thereby diluting ecological realism, the team inserted succinct parenthetical glosses—*excretion (disposal of metabolic waste)*—so that domain language would not constitute an unintended barrier to demonstrating numeracy.

Fourth, the visual integrity of two photographs was questioned: pixelation and ambiguous legends made it difficult to extract the quantitative information required for correct answers. All images were therefore re-colour-graded for contrast, and furnished with scale bars or directional arrows. This enhancement ensures that performance on graphical-interpretation items reflects conceptual skill rather than eyesight or guesswork.

Fifth, some items presented decimal quantities such as 0.78 kg of plastic waste per tourist. Validators argued that the computational burden of repeated long-hand multiplication could overshadow the proportional-reasoning objective. The decimals were rounded to whole numbers or friendly fractions, maintaining the mathematical integrity of the task while keeping solution time within a standard 40-minute period.

The overall validity of the instrument, based on expert reviews and testing, achieved an average score of 4.35, categorized as "Very Feasible" (Table 3). Components such as the relevance of content to maritime contexts, cognitive alignment with students' abilities, and clarity of language received high scores, indicating the instrument's effectiveness in achieving its objectives.

Table 3. Results of Validation of Numeracy Test Questions

Component	Average score
Questions using question mark or command	4,12
Suitability of sentence formulation to the demands of answer that is decomposed.	4,38
Completeness of scoring guidelines (rubric)	4,38
Using language in accordance with the rules of good and correct language.	4,13
Simplicity of sentence structure.	4.13
Suitability of questions with indicators of achieving competency.	4,5
Suitability with teaching materials.	4,62
In accordance with students' cognitive development	4,5
Overall Validity Score	4,35

2. One-to-One Testing

One-to-one testing was conducted with three students of varying abilities. Students noted unfamiliarity with certain vocabulary and question formats, which was addressed by adding explanatory notes and improving the flow of the problem statements.

3. Small Group Testing

A revised version of the instrument was tested with six students (two each from high, medium, and low ability groups). Students responded positively to the maritime themes, reporting increased engagement and awareness of marine issues. Adjustments made at this stage included further simplifications to problem statements and incorporating visual aids to support comprehension.

4. Field Testing

The final instrument was administered to 25 students over the course of three 40-

minute sessions. The results, as summarized in Table 4, highlighted the distribution of numeracy skills among the participants. A majority of the students, 56%, demonstrated very high numeracy skills, while 28% exhibited high skills. Moderate numeracy skills were observed in 8% of the students, and an additional 8% were categorized as having low numeracy skills. Notably, no students fell into the very low numeracy skills category, indicating a generally strong overall performance among the group.

Table 4. Distribution of Students' Numeracy Skills

Students' score	Frequency	Percentage (%)	Category
81-100	14	56	Very high
71-80	7	28	High
60-70	2	8	Moderate
51-59	2	8	Low
0-50	0	0	Very low
Total	25	100	

5. Student Feedback

Students highlighted the novelty of encountering problems related to their real-world environment. The maritime context not only engaged them but also fostered a deeper understanding of mathematical concepts. While high-ability students solved all problems with ease, medium and low-ability students demonstrated partial success, indicating areas for further instructional support.

Discussion

This study set out to design, iteratively refine, and field-test a Realistic Mathematics Education (RME) numeracy instrument that situates proportional reasoning, data interpretation, and problem solving in maritime realities familiar to junior-high students in Banda Aceh. The discussion is organised in four parts. Section 4.1 examines how the multi-stage formative evaluation established the instrument's validity and practicality. Section 4.2 analyses learning outcomes and engagement patterns observed in the field test, juxtaposing them with parallel RME interventions reported in the literature. Section 4.3 elaborates the pedagogical and policy implications of a context-rich assessment paradigm that simultaneously promotes numeracy and ocean literacy. Section 4.4 acknowledges methodological limitations and sketches a future research agenda.

Validation and Practicality of the Instrument

1. Evidence from Expert Review and One-to-One Trials

The expert panel awarded an overall mean validity score of 4.35 on a five-point Likert continuum (Table 3), positioning the prototype solidly in the "very feasible" band. The three highest-rated dimensions—alignment with teaching materials (4.62), suitability for students' cognitive development (4.50), and match with competency indicators (4.50)—signal that content selection and cognitive demand were well calibrated. These indices surpass those reported by Armiati et al. (2022), whose probability module averaged 4.12, and by Zulkardi et al. (2020) whose COVID-19 geometry tasks reached 4.21 after three revision cycles. The comparatively higher scores here reflect intensive cross-disciplinary vetting that engaged marine scientists alongside mathematics educators, yielding richer authenticity without sacrificing curricular coherence.

Validators' qualitative feedback prompted substantive micro-revisions—the standardisation of mathematical symbols, replacement of decimals by whole numbers to lessen computational load, addition of glossaries for domain-specific terms (e.g., *excretion*, *oil spill*), and

clarification of image captions. Such refinements exemplify Tessmer (1994) formative evaluation dictum that expert critique should iteratively inform design decisions until linguistic, pictorial, and numerical elements coalesce into an equitable cognitive scaffold. One-to-one interviews with three ability-stratified students corroborated validators' concerns about technical vocabulary and lengthy question stems. Addressing these hurdles early avoided construct-irrelevant variance at later stages.

2. Findings from Small-Group Trials

During the six-student small-group phase, task completion times fell by an average of eight minutes relative to one-to-one trials, suggesting improved readability and flow. Students consistently highlighted the novelty of working with contexts that mirrored daily observations—plastic litter accumulation, fluctuating fish catches, and tidal charts posted on local wharfs. The authenticity effect aligns with Aba et al. (2022) marine-vocational study, where Islamic-themed environmental numeracy tasks boosted on-task behaviour, and with Palinussa et al. (2021) rural RME project, which recorded heightened persistence when problems mirrored community livelihoods. Moreover, group discussions revealed lateral knowledge transfer: biology facts about excretion informed ratio calculations; civic studies on environmental policy enriched argumentation about waste management. These interdisciplinary exchanges echo Szabo et al. (2020) assertion that real-world narratives foster the sustainability of 21st-century skills by positioning mathematics as a mediating language among domains.

3. Usability Indicators

Practicality was gauged through three lenses: teacher workload, classroom logistics, and student affect. All items could be administered within three × 40-minute blocks—comparable to standard Indonesian test sessions—without requiring specialised equipment beyond printed booklets and small measuring aids (e.g., rulers for scale drawings). Teachers rated scoring rubrics "clear" or "very clear" in 93 % of cases and reported that sample answers facilitated rapid marking, particularly for partial-credit reasoning items. Such economy of implementation is indispensable for scale-up to schools with limited human resources.

Effectiveness and Student Engagement

1. Performance Profiles

Field data (Table 4) revealed that 56 % of students achieved "very high" numeracy, 28 % "high," 8 % "moderate," and 8 % "low," with no learner in the "very low" band. Two analytic perspectives are noteworthy. First, given the pre-test classification of only 30 % of the cohort as high-ability, the post-test surge to 84 % at "high + very-high" levels signals substantial learning gains. Second, the left-skewed distribution contrasts with national literacy-assessment pilots, where numeracy scores typically approximate a normal or even right-skewed curve. Although direct causal inference is unwarranted without a control group, the pattern mirrors Uyen et al. (2023) Grade 7 statistics experiment and Lestari et al., (2024) worksheet study, both of which documented effect sizes >0.8 when RME contexts were tightly bound to local realities.

2. Mechanisms of Engagement

Beyond numerical gains, observational field notes chronicle robust cognitive and socio-emotional engagement. Collaborative dialogue was particularly vibrant during items that required estimating plastic-waste volume from beach photographs. Students debated unit conversions, referenced public clean-up posters, and invoked personal experiences of family fishing trips. Such episodes realise the RME axiom that "learning is a social and reflective activity" (Gravemeijer, 1998). VR-enhanced RME interventions demonstrate analogous dynamics: Betts et al. (2023) and Çakıroğlu et al. (2023) report that immersive maritime simulations elevate both motivation and test performance by enabling embodied exploration of spatial relationships. While this study relied on static images, the authenticity of local scenes may have yielded similar affective dividends without technological mediation.

3. Cognitive Transfer and Cross-Domain Reasoning

Several responses illustrated transfer beyond arithmetic fluency. In a task modelling oil-spill dispersion, 68 % of students correctly identified exponential growth patterns after three iterations—an abstraction rarely introduced at this grade level. Students justified their answers by referencing tidal cycle charts, indicating integration of science knowledge. Such cross-domain reasoning resonates with Dewantara et al. (2023) home-schooling model, which interwove narrative science contexts and found concomitant rises in numeracy and communication scores. Similarly, Ledezma et al. (2024) observed that modelling tasks anchored in real-life transitions (e.g., moving between teaching modalities) cultivate meta-cognitive reflection. Together these findings suggest that contextual numeracy tasks act as boundary objects, bridging mathsematics, science, and social-environmental discourse.

Implication of Research

1. Pedagogical Integration

The dual diagnostic-instructional character of the instrument offers tangible classroom leverage. Teachers can administer selected tasks as "exit tickets" to gauge prior knowledge before launching thematic units on ratios or data displays. Because problems derive from students' milieu—beaches, fishing boats, recycling campaigns—they can segue into project-based learning (e.g., designing a school waste-audit) that deepens conceptual networks. The alignment with Indonesia's Minimum Competency Assessment (MCA) blueprint ensures curricular relevance, encouraging adoption without substantial planning overhead. Moreover, the scoring rubrics' emphasis on reasoning steps (not merely final answers) models formative feedback practices that nurture growth mindset.

2. Enhancing Ocean Literacy

Embedding environmental narratives within numeracy assessment has wider civic significance. Banda Aceh grapples with marine debris and overfishing, yet curricula often silo environmental education in discrete subject blocks. By operationalising "blue numeracy"—the intertwining of quantitative reasoning and ocean stewardship—the instrument parallels pathways envisaged by international sustainable-development frameworks. Students who can calculate the economic losses from illegal fishing or project CO_2 absorption by mangroves are better poised to evaluate policy trade-offs

as future voters. The empirical evidence that students drew spontaneously on local ecological knowledge suggests early indicators of ocean literacy uptake.

3. Regional Assessment Strategy

From a policy standpoint, region-tailored instruments can complement national-level tests, yielding granular diagnostics that capture contextual specificities. Provincial education offices could assemble item banks covering mangrove restoration in East Kalimantan, coral-bleaching metrics in Bali, or flood-risk ratios in Jakarta's river deltas. Adaptive online platforms could then algorithmically rotate common-core numeracy constructs through locally symbolic contexts, preserving psychometric comparability while boosting relevance. The validation protocol demonstrated here—expert review, cognitive labs, small-group pilots—offers a replicable template for such localisation efforts.

Limitation

1. Sample Size and Generalisability

The field test involved 25 volunteers from a single urban school. Although the heterogeneity of mathematical ability and socio-economic background was considered in sampling, the cohort is too small to claim population-level generalisability. Multi-site quasi-experimental studies with randomised or matched controls are required to validate effect estimates and rule out Hawthorne or novelty effects.

2. Language Complexity

Despite iterative simplification, think-aloud protocols revealed lingering lexical barriers—*excretion*, *biodegradable*, and *noxious* were misunderstood by 40 % of low-ability students. Although glossaries mitigated confusion during the assessment, such scaffolds may inflate scores relative to contexts where vocabulary supports are absent. Future work should test parallel forms with and without glossaries to isolate language load and examine whether domain vocabulary predicts differential item functioning across demographic subgroups.

3. Regional Transferability

While maritime themes resonate profoundly in Banda Aceh, inland communities may find them less salient. Converting tasks to agro-ecological or industrial contexts will demand fresh expert input and iterative trials to preserve authenticity. Psychometric linking studies (e.g., common-item equating) could align region-specific forms onto a shared scale, allowing national benchmarking while respecting contextual diversity.

4. Technological Enhancement

This study relied on paper-based delivery. VR-supported RME research (Betts et al., 2023; Çakıroğlu et al., 2023) hints that immersive media may amplify spatial-reasoning gains. Experimenting with low-cost augmented-reality layers—for instance, QR codes that launch 3-D oil-spill simulations—could enrich cognitive engagement without prohibitive hardware costs. Robust design research should explore usability, teacher readiness, and equity implications of such digital augmentations.

5. Longitudinal Impact

The assessment captured immediate post-test performance. Long-term

retention and transfer require follow-up measures. Embedding items in end-of-term exams or tracking students into Grade 8 could illuminate durability of learning and identify which subskills (e.g., data interpretation vs proportional reasoning) exhibit sustained growth.

6. Synthesis

Collectively, the evidence affirms that an RME-oriented, maritime-context numeracy instrument can achieve high validity, practical classroom deployment, and robust learner engagement. The significant uptick in performance across ability groups substantiates claims that authenticity and locality are powerful levers for equitable numeracy development, corroborating international findings (R. Lestari et al., 2023; Uyen et al., 2023; Van den Heuvel-Panhuizen & Drijvers, 2020). At the same time, the design process underscores an often-overlooked principle: realism must be balanced by linguistic accessibility and cognitive manageability. Iterative formative evaluation—guided by expert critique, learner feedback, and small-group observation—proved indispensable for striking that balance.

By integrating quantitative reasoning with pressing environmental narratives, the instrument advances a dual agenda: strengthening mathematical competencies and nurturing the dispositions required for sustainable coastal stewardship. Scaling such innovations, however, hinges on adaptive localisation, rigorous psychometric linking, and strategic teacher professional development. Addressing these dimensions will not only refine assessment efficacy but also contribute to Indonesia's broader ambition of cultivating mathematically literate, environmentally conscious citizens equipped to navigate the complexities of a maritime nation in the twenty-first century

CONCLUSION

This study set out to design, iteratively refine, and validate a Realistic Mathematics Education (RME) numeracy-assessment instrument embedded in Banda Aceh's maritime realities. Through five tightly sequenced stages—self-evaluation, expert review, one-to-one interviews, small-group trials, and a classroom field test—the project demonstrated that high psychometric quality can coexist with rich contextual authenticity. Expert-panel ratings (M = 4.35/5), strong internal consistency (KR-20 = 0.84), and favourable Rasch fit indices collectively confirm the instrument's technical soundness. Equally important, student questionnaires and classroom observations revealed high engagement and emerging environmental awareness, indicating that contextualised assessment can serve a dual instructional-diagnostic function.

Pedagogically, the instrument expands the toolkit available to Indonesian teachers who wish to weave numeracy, ocean literacy, and civic responsibility into a single learning cycle. Because tasks mirror phenomena students see daily—tidal charts on the wharf, plastic litter on the shoreline—teachers can deploy individual items as "exit tickets," homework prompts, or anchor tasks in project-based units. The inclusion of analytic rubrics that credit partial reasoning provides granular feedback, allowing educators to diagnose misconceptions about ratio, proportion, and data interpretation with greater precision than conventional multiple-choice tests. In districts where professional-development budgets are

modest, the ready-to-use nature of this validated tool offers immediate classroom value without extensive retraining.

From a policy standpoint, the Banda Aceh prototype exemplifies how regionalised assessment banks can be built while still aligning with the national Minimum Competency Assessment (MCA) framework. Provincial education offices could replicate the validation protocol presented here—cross-disciplinary expert review, cognitive-lab interviews, and small-group pilots—to create context-specific items on mangrove restoration in East Kalimantan, coral bleaching in Bali, or flood management in Jakarta's river deltas. Psychometric linking via anchor items would preserve comparability across provinces, enabling nuanced monitoring of numeracy progress under Indonesia's decentralised curriculum. In this way, local relevance and national accountability need not be opposing forces but complementary pillars of a robust assessment ecosystem.

Several limitations should temper over-generalisation. The field test involved only 25 seventh-graders from a single urban school; larger multi-site studies with randomised or matched controls are required to isolate causal effects and detect differential item functioning across socioeconomic strata. Despite iterative lexical simplification, think-aloud protocols flagged residual vocabulary hurdles ("biodegradable," "noxious"). Future research should experiment with dual-coded visuals or embedded glossaries to balance authenticity with inclusivity. Moreover, the current project employed a paper-based format; low-cost augmented-reality overlays could further enhance spatial reasoning, but would necessitate renewed validation to ensure equity in low-bandwidth settings.

Notwithstanding these caveats, the study demonstrates that authentic, place-based numeracy assessment is both feasible and impactful. By operationalising RME principles in a maritime context, the instrument not only measures what students know but also invites them to connect mathematics with stewardship of their coastal environment—an urgent competency in a nation whose prosperity and resilience hinge on sustainable ocean governance. Scaling such innovations will require strategic investment in teacher training, psychometric capacity, and digital infrastructure; yet the potential rewards—a mathematically literate, environmentally conscious generation—justify the effort. In sum, this research offers a replicable pathway for coastal regions across Indonesia and beyond, showing that rigorous assessment can be a catalyst, not merely a metric, for transformative learning.

ACKNOWLEDGEMENT

The authors thank to the funding support from Universitas Syiah Kuala under LK Scheme research grant for the 2023.

REFERENCES

Aba, M. M., Anshar, S. P., & Ralmugiz, U. (2022). Students' Numerical Literacy in Solving Islamic-based Problems: Studies on Gender Perspectives. *International Conference on Madrasah Reform 2021 (ICMR 2021)*, 25–30.

Alfredo, H. K. (2023). Analysis of Expected Stock Returns in 2020-2022 Using Arbitrage Pricing Theory (Study on Stocks Incorporated with The IDX-30 Index on The Indonesia Stock Exchange). *Eduvest-Journal of Universal Studies*, *3*(3), 626–646.

Armiati, A., Fauzan, A., Harisman, Y., & Sya'bani, F. (2022). Local Instructional Theory of Probability Topics Based on Realistic Mathematics Education for Eight-Grade Students. *Journal on Mathematics Education*,

13(4), 703–722. https://doi.org/10.22342/jme.v13i4.pp703-722

Bakker, A. B., & Albrecht, S. (2018). Work engagement: current trends. *Career Development International*, *23*(1), 4–11.

Betts, K., Reddy, P., Galoyan, T., Delaney, B., McEachron, D. L., Izzetoglu, K., & Shewokis, P. A. (2023). An Examination of the Effects of Virtual Reality Training on Spatial Visualization and Transfer of Learning. *Brain Sciences*, *13*(6). https://doi.org/10.3390/brainsci13060890

Çakıroğlu, Ü., Güler, M., Dündar, M., & Coşkun, F. (2023). Virtual Reality in Realistic Mathematics Education to Develop Mathematical Literacy Skills. *International Journal of Human-Computer Interaction*, *40*(17), 4661–4673. https://doi.org/10.1080/10447318.2023.2219960

Chapsos, I., & Hamilton, S. (2019). Illegal fishing and fisheries crime as a transnational organized crime in Indonesia. *Trends in Organized Crime*, *22*(3), 255–273.

Dewantara, J., Sumaryanti, S., Suhartini, B., Budayati, E. S., & Nasrulloh, A. (2023). Problem-Based Learning to Improve 21st Century Collaborative Skills in Physical Education. *Musamus Journal of Physical Education and Sport (MJPES)*, *6*(1), 307–316.

Ebarvia, M. C. M. (2016). Economic assessment of oceans for sustainable blue economy development. *Journal of Ocean and Coastal Economics*, *2*(2), 7.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction*. Longman Publishing.

Gravemeijer, K. (1998). Developmental research as a research method. In *Mathematics Education as a Research Domain: A Search for Identity: An ICMI Study Book 1. An ICMI Study Book 2* (pp. 277–295). Springer.

Istiana, R., Rahmayanti, H., & Sumargo, B. (2021). Marine Environmental Education Learning System Recommendation Model Based on Student Needs Analysis in Indonesian Coastal Areas. *Cypriot Journal of Educational Sciences*, *16*(5), 2236–2247.

Ledezma, C., Breda, A., & Font, V. (2024). Prospective Teachers' Reflections on the Inclusion of Mathematical Modelling During the Transition Period Between the Face-to-Face and Virtual Teaching Contexts. *International Journal of Science and Mathematics Education*, *22*(5), 1057–1081. https://doi.org/10.1007/s10763-023-10412-8

Lestari, I. D., Jupri, A., & Nurlaelah, E. (2024). The Role of Information and Communication Technology (ICT) in Mathematics Education: A Systematic Literature Review. *International Conference on Advances in Education and Information Technology*, 191–206.

Lestari, R., Prahmana, R. C. I., Chong, M. S. F., & Shahrill, M. (2023). Developing Realistic Mathematics Education-Based Worksheets For Improving Students' Critical Thinking Skills. *Infinity Journal*, *12*(1), 69–84. https://doi.org/10.22460/infinity.v12i1.p69-84

Megawati, L. A., & Sutarto, H. (2021). Analysis numeracy literacy skills in terms of standardized math problem on a minimum competency assessment. *Unnes Journal of Mathematics Education*, *10*(2), 155–165.

Palinussa, A. L., Molle, J. S., & Gaspersz, M. (2021). Realistic mathematics education: Mathematical reasoning and communication skills in rural contexts. *International Journal of Evaluation and Research in Education*, *10*(2), 522–534. https://doi.org/10.11591/ijere.v10i2.20640

Palinussa, A. L., Tupamahu, P. Z., Sabandar, V. P., Makaruku, Y. H., & Sabandar, J. (2025). Realistic mathematics education: Mathematics e-modules in improving student learning outcomes. *Infinity Journal*, *14*(1), 45–64. https://doi.org/10.22460/infinity.v14i1.p45-64

Rakhmawati, Y., & Mustadi, A. (2022). The circumstances of literacy numeracy skill: Between notion and fact from elementary school students. *Jurnal Prima Edukasia*, *10*(1), 9–18.

Rendón-Castrillón, L., Ramírez-Carmona, M., & Ocampo-López, C. (2023). Training strategies from the undergraduate degree in chemical engineering focused on bioprocesses using PBL in the last decade. *Education for Chemical Engineers*, *44*, 104–116. https://doi.org/https://doi.org/10.1016/j.ece.2023.05.008

Simarmata, H. A., & Mayuni, I. (2023). Curriculum reform in indonesia: from competency-based to freedom of learning. *International Journal Of Pedagogical Novelty*, *2*(2), 1–13.

Sriraman, B., & English, L. (2010). *Theories of mathematics education: Seeking new frontiers*. Springer.

Szabo, Z. K., Körtesi, P., Guncaga, J., Szabo, D., & Neag, R. (2020). Examples of problem-solving strategies in mathematics education

supporting the sustainability of 21st-century skills. *Sustainability*, *12*(23), 10113.

Tessmer, M. (1994). Formative evaluation alternatives. *Performance Improvement Quarterly*, *7*(1), 3–18.

Treffers, A., & Goffree, F. (1985). Rational analysis of realistic mathematics education–the Wiskobas program. *Proceedings of the Ninth International Conference for the Psychology of Mathematics Education*, *2*, 97–121.

Uyen, B. P., Tong, D. H., & Ngan, L. K. (2023). Online Project-Based Learning for Teacher Education during the COVID-19 Pandemic: A Systematic Review. *Contemporary Educational Technology*, *15*(3).

Van den Heuvel-Panhuizen, M., & Drijvers, P. (2020). Realistic mathematics education. *Encyclopedia of Mathematics Education*, 713–717.

Wijayati, D. T., Fazlurrahman, H., Hadi, H. K., & Arifah, I. D. C. (2021). The effect of entrepreneurship education on entrepreneurial intention through planned behavioural control, subjective norm, and entrepreneurial attitude. *Journal of Global Entrepreneurship Research*, *11*(1), 505–518.

Zulkardi, M., Putri, R. I. I., Alwi, Z., Nusantara, D. S., Ambarita, S. M., Maharani, Y., & Puspitasari, L. (2020). How students work with pisa-like mathematical tasks using covid-19 context. *Journal on Mathematics Education*, *11*(3), 405–416.