(JIML) JOURNAL OF INNOVATIVE MATHEMATICS LEARNING
Volume 9, No. 3, September 2026
P-ISSN 2621-4733
E-ISSN 2621-4741
https://dx.doi.org/10.22460/jiml.v9i3.p33259

DEVELOPING UNIMATH: A THORNDIKE-INTEGRATED INTERACTIVE WEB TO IMPROVE HIGH SCHOOL STUDENTS' NUMERICAL ABILITY

Sulthon Aulia Jati[1], Rahayu Kariadinata[2], Rikrik Nurdiansyah[3]
[1]UIN Sunan Gunung Djati, Jl. A.H. Nasution No. 105, Bandung, Indonesia.
sulthonauliajati@gmail.com
[2]UIN Sunan Gunung Djati, Jl. A.H. Nasution No. 105, Bandung, Indonesia.
rahayu.kariadinata@uinsgd.ac.id
[3]UIN Sunan Gunung Djati, Jl. A.H. Nasution No. 105, Bandung, Indonesia.
rikrik@uinsgd.ac.id

ARTICLE INFO

Article history:

Received Jul 15, 2026
Revised Jul 17, 2026
Accepted Jul 31, 2026

Keywords:

Thorndike Theory
Interactive Web
Numerical Ability
4D Development Model
Mathematics Education

ABSTRACT

Numerical ability is essential for interpreting quantitative information, yet Indonesian students often experience difficulty with contextual mathematical problems. Structured web-based practice may address this need when repetition, feedback, and remediation are pedagogically integrated. This study aimed to develop and evaluate UniMath, a Thorndike-integrated web platform for high school students. Research and development followed the 4D model: Define, Design, Develop, and a limited Disseminate stage. Needs data were collected from 34 students and one mathematics teacher, while the main implementation involved 36 Grade X students selected through cluster random sampling. Product feasibility was examined by media and subject-matter experts. Operational change was assessed using a one-group pretest-posttest design, Practice Mode logs, and normalized gain, whereas practicality was evaluated through a 15-item questionnaire and written interviews with six purposively selected students. The media expert judged UniMath feasible after five minor revisions, and the subject-matter expert judged it feasible without revision. Across two practice days, 3,615 attempts were recorded, accuracy increased from 64.02% to 80.64%, the mean score increased from 45.00 to 69.72, and normalized gain reached 0.449 (moderate); practicality was 82.22% (very practical). UniMath was feasible, showed moderate within-class improvement, and was highly practical, although causal inference and broader generalization remain limited.

Corresponding Author:

Sulthon Aulia Jati,
Department of Mathematics Education,
Universitas Islam Negeri Sunan Gunung Djati Bandung,
Jl. A.H. Nasution No. 105, Bandung, Indonesia
Email: author_corresponding@ikipsiliwangi.ac.id

How to Cite:

Jati, S. A., Kariadinata, R., & Nudiansyah, R. (2026). Developing Unimath: A Thorndike-Integrated Interactive Web to Improve High School Students' Numerical Ability. *JIML*, 9(3), 678-692.

INTRODUCTION

Numerical ability enables students to interpret quantitative information, select and apply mathematical procedures, and judge whether a result is reasonable in context. These capacities are central to mathematical literacy because learners must not only calculate but also formulate, employ, and interpret mathematics when solving real-life problems (Organisation for Economic Co-operation and Development [OECD], 2023a). Strengthening numerical ability is therefore important for supporting informed decision-making and meaningful participation in school and everyday situations.

At the global level, the PISA 2022 mathematics framework provides a relevant benchmark for quantitative reasoning, although mathematical literacy and the numerical-ability construct used in this study are not identical. At the national level, Indonesian 15-year-olds obtained a mean mathematics score of 366, compared with the OECD mean of 472, and only 18% attained at least Level 2 proficiency, where students begin to interpret how familiar situations can be represented mathematically (OECD, 2023b). At the local level, the preliminary study at SMAN 1 Dayeuhkolot showed that students experienced difficulty interpreting contextual problems, selecting strategies, calculating accurately, and relating results to the problem context. Among 34 respondents, 82.40% indicated a need for adaptive exercises and 97.10% supported mandatory review after repeated errors. These findings point to a need for sustained contextual practice that provides assistance, consequences, and progress information rather than isolated routine exercises.

Web-based interactive media offer one possible response, but technology does not improve learning automatically. A context-specific meta-analysis found positive effects of digital tools in secondary mathematics and science, particularly when technology complemented other instructional approaches (Hillmayr et al., 2020). A randomized field trial likewise found that online mathematics homework combining timely feedback and hints with teacher support increased standardized achievement compared with existing homework practice (Roschelle et al., 2016). However, a recent review of technology use in secondary mathematics reported uneven evaluation designs and a continuing need for stronger measurement across settings (Darmanova et al., 2025). Effective design should therefore combine timely and informative feedback (Shute, 2008), adaptive assistance or difficulty (Chiotaki et al., 2023), and carefully aligned gamification, whose benefits depend on instructional quality rather than on points or badges alone (Sailer & Homner, 2020; Zainuddin et al., 2020).

Previous development studies demonstrate the potential of digital mathematics products in specific contexts. A contextual mathematics website developed through the 4D model was judged valid and attractive (Baihaqi & Wulantina, 2023), while a web-based electronic worksheet was developed to support high school numeracy in geometry (Pamungkas et al., 2023). QuizWhizzer-based ethnomathematics media were reported as highly valid, highly practical, and effective for learning exercises (Muthmainah et al., 2024). Liveworksheet-based materials and numeracy-oriented algebra resources have also produced positive validity, practicality, or learning outcomes (Az Zahra et al., 2025; Hapsari et al., 2025). Thorndike has also informed mathematics-media development: a non-web Rainbow Ladder medium for Grade IV was reported as valid, practical, and effective (Watoni et al., 2024). Taken together, these studies support the educational potential of websites, worksheets, interactive quizzes, and theory-based media, but they provide less detail on a unified system that connects learning theory to explicit adaptive rules, mandatory remediation, separate practice and testing environments, and auditable process logs.

Thorndike's law of exercise and law of effect were selected as the principal design framework for UniMath because they corresponded to the identified need for repeated practice and meaningful consequences after each response. Within the operational framework adopted in this study, the law of exercise concerns the strengthening of situation-response connections through use, whereas the law of effect concerns how consequences influence those connections (Thorndike, 1911). Applications in mathematics education commonly emphasize repeated exercises and reinforcement (Amsari, 2018). In UniMath, the law of exercise is operationalized through continuing attempts, progression floors, and increasingly challenging tasks. The law of effect is operationalized through immediate correctness feedback, experience points (XP), streaks, staged hints, and a mandatory remedial pathway after repeated errors. Practice Mode is separated from Test Mode so that instructional consequences support learning without directly influencing pretest and posttest performance.

The research gap therefore lies not merely in the availability of another mathematics website, but in the limited evidence on theory-to-system integration that can be inspected through both product features and implementation data. The studies cited above do not jointly document a platform that combines repeated adaptive practice, response-contingent consequences, verified remediation, progression monitoring, separate Practice and Test modes, and distinct evidence for feasibility, instrument quality, operational change, and practicality. UniMath was developed to connect these components within one traceable platform while retaining explicit limits on causal interpretation.

Accordingly, this study aimed to: (1) develop UniMath through the 4D model; (2) determine its media and subject-matter feasibility and document supporting test-instrument evidence; (3) examine within-class operational changes in students' numerical ability after a limited implementation; and (4) determine practicality from student questionnaire and written interview data.

METHOD

Research Design and Participants

This Research and Development study used the 4D model: Define, Design, Develop, and Disseminate (Thiagarajan et al., 1974). Dissemination was deliberately limited to use, orientation, documentation, and packaging at the target school; it did not constitute large-scale distribution. Quantitative and qualitative data were analysed according to their respective functions and integrated at interpretation through convergence matrices and joint displays (Fetters et al., 2013).

The study was conducted at SMAN 1 Dayeuhkolot, Bandung Regency, Indonesia, from December 2025 to July 2026. The sampling frame consisted of ten intact Grade X classes. A one-time computerized random draw selected class X-9, containing 36 students, as the implementation cluster. Reporting the cluster frame and random-selection procedure was intended to preserve methodological traceability (Ballance, 2024). Two students were absent because of illness during the initial needs questionnaire, leaving 34 valid needs responses; all 36 students participated in the main implementation. One Grade X mathematics teacher served as a key informant during the Define stage only.

Formal institutional permission to conduct the study and collect the required school-based data was obtained through the university and the participating school before data collection. Student participation took place under the school's authorization, and student identities were anonymized in all analytical outputs.

Operational change in numerical ability was examined with a pre-experimental one-group pretest-posttest design. The design compared the same students before and after the intervention but did not randomly assign individuals to treatment and control groups. Accordingly, the analysis was restricted to within-class change and did not claim that all observed change was caused exclusively by UniMath (Campbell & Stanley, 1963).

Table 1. Four-D development activities and traceable outputs

Stage	Principal activities	Output/evidence
Define	Preliminary study; student-teacher needs; readiness and cluster selection	Converged needs and implementation conditions
Design	Content, adaptive rules, Practice/Test modes, instruments, and analytics plan	Product and measurement specifications
Develop	Programming, technical test, expert appraisal, test judgement/CTT, revision, and implementation	Functional revised product and layered evidence
Limited Disseminate	Target-school use, orientation, guidance, documentation, and packaging	Context-bound adoption evidence

Product and Intervention

UniMath is an interactive web platform with Student, Teacher, and Admin roles. Its main content comprises six contextual social-arithmetic units: discounts and final prices; value-added tax, service fees, and delivery costs; profit, loss, and percentages; simple interest; gross, net, and tare; and promotions and transaction decisions. Three remedial units address basic percentage operations, total payment and currency operations, and division and unit prices. The Practice Mode bank contains 450 multiple-choice items across 50 progression floors and 72 comprehension checkpoints.

Each practice session begins at medium difficulty. A correct answer increases the progression floor by one, adds XP and a streak, and routes the next item to a harder level. The first incorrect response resets the streak, keeps the current floor, routes the next item to an easier level, and makes Hint 1 available. A second consecutive error activates Hint 2. A third consecutive error activates Wajib Belajar, a mandatory remedial pathway. Students may return to practice at the same floor after viewing at least 80% of the material, meeting a minimum reading time, and answering the checkpoint with 100% accuracy.

Test Mode contains separate ten-item pretest and posttest forms with five options per item. It provides no XP, streak, badge, achievement, confetti, hint, or floor progression. The pretest was administered on 8 June 2026; Practice Mode on 9 and 10 June; the posttest on 11 June; and the practicality questionnaire and follow-up written interviews on 13 June.

Table 2. Adaptive Practice Mode mechanism and theoretical function

Condition	System response	Instructional function
Session start	Medium-difficulty item	Balanced entry challenge
Correct	Floor +1; XP/streak +; harder next item	Exercise, reinforcement, challenge
First consecutive error	Same floor; streak reset; easier next item; Hint 1	Initial recovery
Second consecutive error	Hint 2	More explicit assistance
Third consecutive error	Mandatory remedial pathway	Prerequisite recovery

Return requirement	80% viewing; minimum time; checkpoint 100%	Verified remedial engagement
Return to practice	Resume at the same floor	Retry without punitive regression

Instruments and Data Collection

Data sources included needs questionnaires, teacher responses, preliminary observation, media and subject-matter expert-appraisal forms, functional-trial and implementation observations, Practice Mode logs, pretest-posttest responses, a 15-item four-point practicality questionnaire, follow-up written interviews, and development documentation. The four numerical-ability indicators used consistently in scoring were mathematical calculation, logical reasoning, problem solving, and recognition of numerical patterns and number relationships, aligned with the study's operational framework for numerical reasoning (Hidayatullah et al., 2024; OECD, 2023a).

Media and subject-matter appraisals were conducted separately because they addressed different objects. The media expert examined interface, readability, navigation, interactivity, responsiveness, technical functions, the mini-LMS, gamification, and feature flow through direct web access. The subject-matter expert examined the six main and three remedial units, the 450-item map, 50 representative practice items, 72 checkpoints, hints, explanations, and remedial direction. Expert decisions and comments were retained in their original qualitative form rather than converted into invented percentages.

Twenty final pretest-posttest items underwent theoretical judgement for content representation, construction, language, cognitive demand, answer keys, distractors, and planned pair equivalence. The ten-item pretest form was then tried empirically with 30 non-target respondents and analysed through classical test theory (CTT). Content evidence and empirical item behaviour were treated as complementary rather than interchangeable, consistent with educational-testing standards (American Educational Research Association et al., 2014). The posttest was not claimed to have the same empirical CTT evidence; its support came from judgement, blueprint alignment, and planned pair equivalence. Item-total correlation, KR-20, difficulty, discrimination, and distractor function were examined using procedures consistent with contemporary item-analysis guidance (Rashwan et al., 2024; Rezigalla et al., 2024).

Data Analysis

Closed needs and practicality responses were summarized with frequencies and percentages. Two negatively worded practicality items were reverse scored. Pretest and posttest scores were summarized descriptively, and class normalized gain was calculated as (posttest mean - pretest mean)/(100 - pretest mean). Values below 0.30 were interpreted as low, 0.30 to below 0.70 as moderate, and 0.70 or higher as high (Hake, 1998). Practice logs were used only to describe exercise, system consequences, and implementation processes; they were not used as the outcome measure of effectiveness.

Open responses and six follow-up written interviews were analysed thematically through familiarization, initial coding, theme development, review, definition, and reporting (Braun & Clarke, 2006). The six cases were selected through stratified purposive sampling: Day 1 XP, accuracy, and attempts were used to form upper, middle, and lower activity-performance strata, and two informative cases were selected from each stratum (Palinkas et al., 2015). Student identities were anonymized in analytical outputs. Questionnaire and interview findings were then integrated in a joint display.

RESULTS AND DISCUSSION

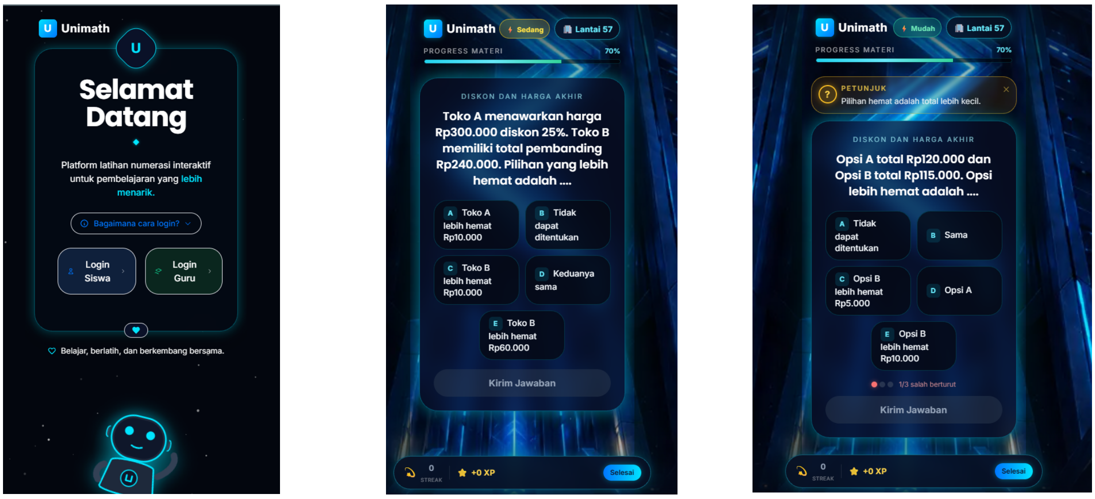
Results

The results are presented in the same order as the four study objectives: the development of UniMath, product feasibility and supporting instrument evidence, within-class operational changes in numerical ability, and practicality.

Development of UniMath

All planned 4D activities were completed with documentary evidence. During Define, the preliminary diagnostic study combined 34 student needs responses, one mathematics-teacher key-informant response, and a readiness observation. Students reported needs for step-by-step guidance (61.80%), worked explanations (64.70%), summaries (73.50%), adaptive exercises (82.40%), mandatory review after repeated errors (97.10%), and progress information (at least 76.50%). All respondents had access to a personal smartphone, although network quality varied. The student and teacher findings converged on the need for contextual problems, staged assistance, remediation, adaptation, and progress monitoring.

During Design, these needs were translated into separate Practice and Test modes, six social-arithmetic units, three prerequisite-remedial units, adaptive rules, staged hints, progress tracking, and restrained gamification. Develop produced a functional web platform, technical testing, expert appraisal, theoretical test judgement, empirical pretest analysis, product revision, a functional trial, and the main implementation. Limited Disseminate introduced the platform and its monitoring functions to the target class and mathematics teacher, provided operating guidance, and packaged the product and its development evidence. All 29 planned development activities were completed; this result indicates process traceability rather than perfection of every product outcome.



(a) User entry *(b) Adaptive practice* *(c) Staged hint*

Figure 1. Representative UniMath interfaces: localized user entry, adaptive Practice Mode, and staged assistance.

Product Feasibility and Supporting Instrument Evidence

The media expert's final decision was "feasible for use with minor revisions." Five revisions were requested: increasing hint-icon visibility, adding controllable background music and volume, providing adjustable text size, improving desktop-mobile proportions, and adding concise guidance from registration through all principal features. All five revisions were completed and checked before the main implementation.

The subject-matter expert judged UniMath "feasible for use without revision" after reviewing the six main units, three remedial units, item map, 50 representative practice items, 72 checkpoints, hints, explanations, and remedial directions. Expert decisions and comments were retained in their original qualitative form rather than converted into percentages that were not produced by the appraisal instruments.

The 20-item pretest-posttest package was judged theoretically suitable in terms of content representation, construction, language, cognitive demand, answer keys, distractors, and planned pair equivalence. In the empirical trial involving 30 non-target respondents, all ten pretest items had corrected item-total correlations above the critical value of 0.361, and KR-20 was 0.937. Three items were easy and seven were moderate, while discrimination ranged from 0.625 to 1.000. Four items had fully functioning distractors at the initial analysis, whereas six required minor distractor revisions. This empirical evidence applies only to the pretest form; support for the posttest was based on theoretical judgement, blueprint alignment, and planned pair equivalence.

Table 3. Feasibility decisions and supporting test-instrument evidence

Evidence source	Result	Decision or boundary
Media expert	Feasible with five minor revisions; all completed	Feasible after revision
Subject-matter expert	M1A-M1F, R1-R3, item map, 50 samples, and 72 checkpoints reviewed	Feasible without revision
Theoretical judgement	Twenty pretest-posttest items reviewed	Package suitable by judgement
Pretest CTT (n = 30)	10/10 corrected item-total criteria; KR-20 = 0.937; discrimination 0.625-1.000	Evidence applies only to pretest

Practice Mode Implementation and Operational Change in Numerical Ability

Across two Practice Mode days, all 36 students generated 3,615 attempts. Attempts increased from 1,673 on Day 1 to 1,942 on Day 2, while correct responses increased from 1,071 to 1,566 and incorrect responses decreased from 602 to 376. Accuracy increased from 64.02% to 80.64%. Visible hints decreased from 667 to 369, total XP increased from 12,340 to 21,110, and the number of students reaching progression floor 20 increased from 17 to all 36 students. These values describe participation, practice processes, and progression; they were not used as the outcome measure of effectiveness.

The logs also recorded the consequences associated with student responses. All 2,637 correct responses received positive XP, whereas all 978 incorrect responses received zero XP. Among auditable transitions, 1,986 correct responses were followed by a one-floor increase, 942 incorrect responses retained the same floor, and 869 of those incorrect transitions (92.25%) recorded a positive hint value. These log patterns were treated as process evidence rather than as an additional outcome measure.

Table 4. Practice Mode activity on Day 1 and Day 2

Indicator	Day 1	Day 2	Change
Students	36	36	0
Attempts	1,673	1,942	+269

Correct responses	1,071	1,566	+495
Incorrect responses	602	376	-226
Accuracy	64.02%	80.64%	+16.62 pp
Visible hints	667	369	-298
Total XP	12,340	21,110	+8,770
Reached floor 20	17	36	+19

The effectiveness dataset contained 36 complete pretest-posttest pairs, comprising 360 pretest responses and 360 posttest responses. Eight non-target or testing accounts were excluded using the authoritative class roster, and no response was imputed or altered. The mean increased by 24.72 points, from 45.00 (SD = 8.78) on the pretest to 69.72 (SD = 11.58) on the posttest. All 36 students improved: 19 students increased by 20 points and 17 increased by 30 points. The class normalized gain was 0.449, categorized as moderate.

Table 5. Descriptive pretest-posttest results for 36 complete pairs

Statistic	**Pretest**	**Posttest**	**Difference**
Mean	45.00	69.72	24.72
Median	45.00	70.00	20.00
Sample SD	8.78	11.58	5.06
Minimum	30.00	50.00	20.00
Maximum	70.00	90.00	30.00
Class normalized gain	-	-	0.449 (moderate)

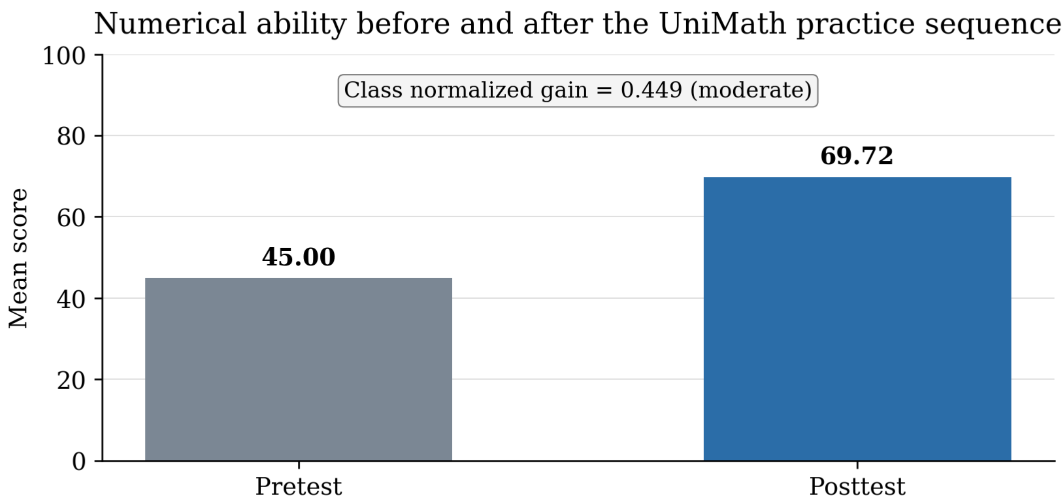


Figure 2. Mean numerical-ability scores before and after the UniMath practice sequence.

Improvement differed across the four indicators. Mathematical calculation increased from 45.37% to 75.00%, with a normalized gain of 0.542; logical reasoning increased from 53.70% to 75.00%, with a gain of 0.460; and problem solving increased from 40.28% to 70.83%, with a gain of 0.512. These three gains were moderate. Numerical-pattern and number-relationship recognition increased from 36.11% to 52.78%, but its normalized gain was 0.261 and categorized as low. Because each indicator was represented by only two or three items, these indicator-level findings were treated as diagnostic directions rather than independently reliable subscales.

Table 6. Numerical-ability achievement by indicator

Indicator	Pretest	Posttest	Gain	Category
I1 Mathematical calculation	45.37%	75.00%	0.542	Moderate
I2 Logical reasoning	53.70%	75.00%	0.460	Moderate
I3 Problem solving	40.28%	70.83%	0.512	Moderate
I4 Numerical patterns/relationships	36.11%	52.78%	0.261	Low

Practicality and User Experience

All 36 students completed the 15-item practicality questionnaire after the pretest, two Practice Mode days, and posttest. Overall practicality was 82.22%, categorized as very practical. The usability and technical aspect reached 82.29%, while the affective-cognitive aspect reached 82.14%. Intention to reuse was the strongest measured aspect at 87.50%, whereas system stability reached 75.00%. The item concerning excessive orientation toward gamified rewards was the lowest at 68.06%, indicating that positive motivation coexisted with a risk of focusing excessively on points or progression.

The six follow-up written interviews generated five cross-case themes. First, platform use became easier after familiarization. Second, adaptive difficulty and staged assistance supported persistence. Third, gamification encouraged continued practice but required conceptual control. Fourth, practice encouraged more careful and ordered strategies. Fifth, improvement needs differed across the upper, middle, and lower activity-performance profiles. Integration of the questionnaire and interview findings primarily showed confirmation and expansion: the interviews supported the quantitative pattern while explaining why system stability, language clarity, initial orientation, and more explicit hints remained important.

Table 7. Integrated practicality findings

Aspect	Quantitative result	Qualitative explanation	Development implication
Overall practicality	82.22%; very practical	Positive acceptance across profiles	Retain core flow
Usability/technical	82.29%	Easy after familiarization	Short initial orientation
Affective-cognitive	82.14%	Motivation and more ordered strategies	Maintain reflective feedback
System stability	75.00%	Some lag/reset experiences	Improve loading and progress saving
Reward orientation	68.06%	Risk of chasing points	Tie rewards to accuracy and reflection
Intention to reuse	87.50%	High willingness with varied support needs	Differentiate assistance/challenge

Discussions

Development and Theory-to-System Integration

The completion of all 29 planned activities indicates that the development process was traceable across Define, Design, Develop, and limited Disseminate. This result is comparable to Baihaqi and Wulantina (2023), who also employed the 4D model to develop contextual website-based mathematics media and reported that their product was valid and attractive. The distinction is that UniMath did not stop at presenting contextual content: identified needs were translated into explicit adaptive rules, separate Practice and Test modes, mandatory remediation, and process logs. Thus, the contribution lies in connecting needs analysis, theory, system behaviour, and differentiated evidence rather than merely producing another website.

UniMath also extends earlier mathematics-media applications of Thorndike's theory. Watoni et al. (2024), for example, developed a physical Rainbow Ladder medium based on Thorndike's theory and reported that it was valid, practical, and effective in an elementary-school context. UniMath differs in educational level, content, medium, and implementation: the laws of exercise and effect were translated into digital response-contingent rules that could be inspected through repeated attempts, correctness consequences, XP, progression-floor changes, staged hints, and mandatory remediation. These differences do not establish superiority, but they show how theory can be made operational and auditable within a web platform.

The Practice Mode logs support this theory-to-system interpretation. Repeated attempts and continued progression corresponded to the operational use of the law of exercise, while positive XP after correct responses and zero XP, retained floors, hints, or remediation after incorrect responses represented different consequences under the law of effect (Thorndike, 1911). Staged hints also corresponded to the principle that formative feedback should provide timely information after learner action (Shute, 2008). Separating Practice Mode from Test Mode prevented XP, hints, streaks, and progression from directly influencing the pretest and posttest scores. The log evidence should nevertheless be interpreted as implementation evidence, not proof that Thorndike's principles alone caused the score change.

Product Feasibility and Instrument Evidence

The expert findings indicate that UniMath was feasible after completing five media revisions and required no subject-matter revision. This direction is consistent with Muthmainah et al. (2024), whose QuizWhizzer-based ethnomathematics learning medium was reported as highly valid, highly practical, and effective. Both studies support the feasibility of interactive digital mathematics exercises, but their development models, content, participants, appraisal instruments, and effectiveness criteria differ. UniMath therefore retained the experts' original qualitative decisions rather than converting them into percentages solely to facilitate comparison.

The instrument findings require a layered interpretation. The theoretical judgement applied to the complete 20-item pretest-posttest package, whereas the empirical CTT analysis involved only the ten-item pretest form. The KR-20 value of 0.937, corrected item-total results, difficulty distribution, discrimination indices, and distractor analysis provide strong supporting evidence for that pretest form, but they cannot be transferred automatically to the posttest. This boundary strengthens the accuracy of the report because it distinguishes evidence that was actually obtained from evidence supported through blueprint alignment and planned pair equivalence.

Operational Change in Numerical Ability

The increase from a mean of 45.00 to 69.72 and the normalized gain of 0.449 indicate moderate within-class improvement after the UniMath practice sequence. The direction of this result is compatible with Hillmayr et al. (2020), whose meta-analysis found an overall positive effect of digital tools in secondary mathematics and science and noted that technology was more beneficial when integrated with instruction rather than treated as a substitute for it. UniMath was likewise implemented within a school-learning context rather than positioned as a replacement for the teacher. However, the meta-analysis synthesized controlled comparisons, whereas the present study used one intact class without a control group; its effect magnitude must therefore not be directly equated with the meta-analytic result.

The result is also directionally consistent with Roschelle et al. (2016), who reported improved standardized mathematics achievement from online homework combining timely feedback, hints, teacher support, and organized information. Their study used a randomized field trial with a control group, whereas UniMath used a shorter one-group pretest-posttest implementation and a locally developed numerical-ability instrument. Consequently, the present increase supports within-class operational improvement but is not an equivalent causal estimate.

Positive numeracy outcomes have also been reported for other digital or structured products. Az Zahra et al. (2025) developed a Liveworksheet-based electronic worksheet using problem-based learning for Grade X numeracy, while Hapsari et al. (2025) developed numeracy-oriented algebra materials for vocational high school students. These studies and UniMath share an emphasis on contextual numerical activity and structured learning materials. They differ, however, in mathematical content, product architecture, development model, duration, assessment instruments, and research design. The similarities support the educational relevance of contextual numeracy resources, but the reported outcomes should not be compared as though they arose from equivalent interventions.

The indicator results provide an additional diagnostic contribution. Moderate gains in mathematical calculation, logical reasoning, and problem solving suggest that these aspects responded more strongly during the limited implementation. In contrast, the low gain of 0.261 in numerical-pattern and number-relationship recognition identifies a specific area for further development. Because each indicator contained only two or three items, this difference should guide item and feature revision rather than be interpreted as a stable ranking of four independent subscales.

Practicality and User Experience

The overall practicality result of 82.22% and the interview themes indicate generally positive acceptance across different activity-performance profiles. This finding is directionally comparable with Muthmainah et al. (2024), who also reported high practicality for an interactive mathematics exercise medium. UniMath adds explanatory qualitative evidence: students described easier use after familiarization, support from adaptive difficulty and staged hints, and more careful and ordered practice strategies. At the same time, some students reported lag or reset experiences and different assistance needs, showing that high overall practicality did not mean that every aspect was optimal.

The practicality findings also illustrate the dual role of gamification. Intention to reuse was high at 87.50%, but reward orientation was the lowest item at 68.06%. This pattern is consistent with Sailer and Homner's (2020) meta-analysis, which found significant but generally small effects of gamification on cognitive, motivational, and behavioural outcomes. It also aligns with Zainuddin et al. (2020), who reported positive themes involving engagement and motivation, achievement, and social connection while identifying contradictions and weak theoretical foundations in parts of the literature. UniMath therefore used restrained gamification and linked rewards to correctness and progression, but the results still indicate a need to prevent points from displacing conceptual reflection.

Integrated Contribution, Limitations, and Implications

UniMath's principal contribution lies in connecting needs, Thorndike's principles, system rules, and differentiated evidence. Needs findings were mapped to product features; the laws of exercise and effect were mapped to auditable attempts and response consequences; Practice Mode was separated from Test Mode; and development, feasibility, instrument evidence, implementation processes, operational change, and practicality were examined using distinct data sources. The mandatory remedial pathway also prevented repeated incorrect responses from being treated merely as unsuccessful game attempts.

Several limitations bound this interpretation. The implementation involved one class at one school over a short period and did not include a comparison group. Empirical CTT evidence applied only to the pretest form, while each numerical-ability indicator was represented by only two or three items. Some digital-log structures, including session identifiers and remedial records, were not fully uniform, and dissemination remained limited to the target-school context. These limitations restrict causal language, broader generalization, indicator-level interpretation, and claims about posttest psychometrics; they do not invalidate the documented development process or the within-class findings.

Future development should strengthen tasks involving numerical patterns and relationships, improve loading and progress saving, refine wording and initial orientation, and connect rewards more explicitly to accuracy and reflection. Future research should employ longer implementation periods, multiple schools, a comparison group or stronger experimental design, empirical evaluation of the posttest form, and more consistent session and remediation logs.

CONCLUSION

Based on results and discussions above, we can conclude some points on the following list below:

1. UniMath was developed through the Define, Design, Develop, and limited Disseminate stages of the 4D model as an interactive social-arithmetic practice platform integrating adaptive tasks, immediate feedback, staged hints, mandatory remediation, restrained gamification, and progress monitoring. All 29 planned development activities were completed with traceable evidence.
2. The product was feasible for use after the media expert's five minor revisions were completed, while the subject-matter expert required no revision. The pretest form demonstrated strong supporting empirical evidence, including a KR-20 coefficient of 0.937, although this evidence was not extended to the posttest form.
3. During the limited implementation, the mean numerical-ability score increased from 45.00 to 69.72, producing a class normalized gain of 0.449, categorized as moderate. All 36 students improved, although numerical-pattern and number-relationship recognition remained the weakest indicator. This finding indicates moderate within-class improvement rather than exclusive causal effectiveness.
4. UniMath achieved an overall practicality score of 82.22%, categorized as very practical. Students valued its adaptive assistance, feedback, progression, and retry opportunities, although system stability, wording, initial orientation, and reward-focused behaviour remain priorities for development.

Future research should examine UniMath across multiple schools, larger samples, and longer implementation periods using a comparison group or stronger experimental design. Further studies should empirically evaluate the posttest form, strengthen numerical-pattern and relationship tasks, improve loading and progress-saving reliability, and standardize session and remedial-log structures.

ACKNOWLEDGMENTS

The author acknowledges the research supervisors, media and subject-matter experts, the mathematics teacher, the school leadership, and the Grade X-9 students of SMAN 1 Dayeuhkolot for their contributions to the development and limited implementation of UniMath. No external research funding was reported.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Amsari, D. (2018). Implikasi teori belajar E. Thorndike (behavioristik) dalam pembelajaran matematika. *Jurnal Basicedu, 2(2), 52–60.* https://doi.org/10.31004/basicedu.v2i2.168

Az Zahra, P., Winarni, S., & Rohati. (2025). Pengembangan E-LKPD berbasis Liveworksheet menggunakan pendekatan berbasis masalah dalam meningkatkan kemampuan numerasi kelas X SMA. *Eksponen, 15*(2), 57–68. https://doi.org/10.47637/eksponen.v15i2.1922

Baihaqi, R., & Wulantina, E. (2023). Pengembangan media pembelajaran matematika berbasis website dengan pendekatan kontekstual. *Jurnal Derivat: Jurnal Matematika dan Pendidikan Matematika, 10*(3), 117–128. https://doi.org/10.31316/jderivat.v10i3.4986

Ballance, O. J. (2024). Sampling and randomisation in experimental and quasi-experimental CALL studies: Issues and recommendations for design, reporting, review, and interpretation. *ReCALL, 36*(1), 58–71. https://doi.org/10.1017/S0958344023000162

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Rand McNally.

Chiotaki, D., Poulopoulos, V., & Karpouzis, K. (2023). Adaptive game-based learning in education: A systematic review. *Frontiers in Computer Science, 5*, Article 1062350. https://doi.org/10.3389/fcomp.2023.1062350

Darmanova, Z., Abylkassymova, A., & Nurmukhamedova, Z. (2025). A systematic review of technology use in middle and high school mathematics education: Insights from contextual, methodological, and evaluation characteristics. *Frontiers in Education, 10*, Article 1644284. https://doi.org/10.3389/feduc.2025.1644284

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs: Principles and practices. *Health Services Research, 48*(6 Pt 2), 2134–2156. https://doi.org/10.1111/1475-6773.12117

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64–74. https://doi.org/10.1119/1.18809

Hapsari, T., Rif'at, M., & Kusumastuti, N. (2025). Enhancing numeracy literacy through algebra learning materials: A vocational high school context. *Al-Jabar: Jurnal Pendidikan Matematika, 16*(1), 409–425. https://doi.org/10.24042/ajpm.v16i1.27935

Hidayatullah, M. S., Dewi, R. S., Thaibah, H., Larasaty, I., Simanjuntak, R., & Anggraini, M. (2024). *Mengukur dan memahami penalaran: Teori dan praktik*. Komojoyo Press. https://repo-dosen.ulm.ac.id/bitstream/handle/123456789/33945/Mengukur%20dan%20memahami%20penalaran.pdf

Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S. I., & Reiss, K. M. (2020). The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Computers & Education, 153*, Article 103897. https://doi.org/10.1016/j.compedu.2020.103897

Muthmainah, A. S. N., Jihad, A., & Nurdiansyah, R. (2024). Development of learning media using QuizWhizzer based on ethnomathematics as a learning exercise. *Jurnal Gantang, 9(2), 135–146.* https://doi.org/10.31629/jg.v9i2.6921

Organisation for Economic Co-operation and Development. (2023a). *PISA 2022 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/dfe0bf9c-en

Organisation for Economic Co-operation and Development. (2023b). *PISA 2022 results (Volume I and II): Country notes - Indonesia*. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/11/pisa-2022-results-volume-i-and-ii-country-notes_2fca04b9/indonesia_0e09c072/c2e1ae0e-en.pdf

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health and Mental Health Services Research, 42*(5), 533–544. https://doi.org/10.1007/s10488-013-0528-y

Pamungkas, A. S., Khaerunnisa, E., & Fajriyanti, T. (2023). Development of web-based e-LKPD to develop high school students' numeracy on geometry. *Union: Jurnal Ilmiah Pendidikan Matematika, 11*(3), 538–550. https://doi.org/10.30738/union.v11i3.16091

Rashwan, N. I., Aref, S. R., Nayel, O. A., & Rizk, M. H. (2024). Postexamination item analysis of undergraduate pediatric multiple-choice questions exam: Implications for developing a validated question bank. *BMC Medical Education, 24*, Article 168. https://doi.org/10.1186/s12909-024-05153-3

Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., Alghamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. K. (2024). Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education, 24*, Article 445. https://doi.org/10.1186/s12909-024-05433-y

Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open, 2*(4), Article 2332858416673968. https://doi.org/10.1177/2332858416673968

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review, 32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Thiagarajan, S., Semmel, D. S., & Semmel, M. I. (1974). *Instructional development for training teachers of exceptional children: A sourcebook*. Leadership Training Institute/Special Education, University of Minnesota. https://files.eric.ed.gov/fulltext/ED090725.pdf

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Macmillan. https://archive.org/details/animalintelligen00thor

Watoni, T. M. S., Turmuzi, M., Fauzi, A., & Makki, M. (2024). Pengembangan media tangga pelangi berbasis teori Thorndike terhadap prestasi belajar siswa dalam pembelajaran matematika kelas IV di SDN 01 Gelanggang. *Pendas: Jurnal Ilmiah Pendidikan Dasar, 9*(1). https://doi.org/10.23969/jp.v9i1.12681

Zainuddin, Z., Chu, S. K. W., Shujahat, M., & Perera, C. J. (2020). The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review, 30*, Article 100326. https://doi.org/10.1016/j.edurev.2020.100326