PAEDAGOGIA
Jurnal Penelitian Pendidikan
http://jurnal.uns.ac.id/paedagogia

DOI : 10.20961/Paedagogia.v29i3.127667
Vol. 29, No. 3, 2026, pp. 399-415
p-ISSN 0126-4109
e-ISSN 2549-6670

Museum Digital Ruang-Geo and Junior Secondary Students' Geometric Ability on Academic Competency Test (TKA)-Aligned Tasks: A Development and Quasi-Experimental Study

Sigit Rimbatmojo[1*], Eka Farida Fasha[1], Khurotul Aeni[2]
[1] Mathematics Education Study Program, Faculty of Teacher Training and Education, Universitas Peradaban, Brebes, Indonesia.
[2] Informatics Study Program, Faculty of Science and Technology, Universitas Peradaban, Brebes, Indonesia.

Keywords: *Digital Museum; Geometric Ability; Interactive 3D Visualisation; Quasi-Experiment; TKA-Aligned Tasks*

Article history
Received: 28 August 2026
Revised: 9 October 2026
Accepted: 25 October 2026
Published: 31 October 2026

*Corresponding Author Email:
sigitrimbaatmojo@gmail.com

DOI: 10.20961/paedagogia.v29i3.127667

© 2026 The Authors. This open-access article is distributed under a CC BY-SA 4.0 DEED License

Abstract: Geometry tasks in the Academic Competency Test (Tes Kemampuan Akademik, TKA) for junior secondary students demand visualisation and reasoning, yet in the school studied 85.94% of students reported limited learning media and 78.13% reported difficulty with TKA-type reasoning. This study developed Museum Digital Ruang-Geo, a mobile virtual-gallery application whose object pages combine interactive three-dimensional (3D) models, nets, formulae and TKA-aligned practice, and examined its expert-rated quality and preliminary learning outcomes. A sequential design combined a descriptive needs analysis, ADDIE-based development with expert and practitioner validation, and a quasi-experiment with two intact Grade IX classes (n = 32 each) comparing Ruang-Geo with worksheets and PowerPoint; the posttest was taken in the application by one class and on paper by the other. Mean validation scores (1–5) were 4.4 (content experts, n = 2), 4.5 (media experts, n = 2), and 4.75 (one practitioner). Both classes improved significantly (Ruang-Geo: 52.97 to 70.16, dz = 2.47; control: 51.72 to 68.13, dz = 1.94; both p < .001). Between-group differences were not significant for posttest scores ($t(62)$ = 0.91, p = .367, 95% CI [−2.44, 6.50]), N-Gain (0.371 vs 0.326; p = .109, d = 0.41) or the baseline-adjusted comparison (p = .326). The novelty lies in organising 2D and 3D geometry as explorable exhibits traced to TKA indicators; the study offers preliminary evidence of validity, feasibility and learning gains, but not of superiority over worksheets and PowerPoint.

How to cite: Rimbatmojo, S., Fasha, E. F., & Aeni, K. (2026). Museum Digital Ruang-Geo and Junior Secondary Students' Geometric Ability on Academic Competency Test (TKA)-Aligned Tasks: A Development and Quasi-Experimental Study. *PAEDAGOGIA, 29*(3), 399-415. DOI: 10.20961/paedagogia.v29i3.127667

INTRODUCTION

Geometry is a core strand of junior secondary mathematics, and it is also a context in which students develop visualisation, spatial thinking and reasoning. These capacities are relevant to the Academic Competency Test (Tes Kemampuan Akademik, TKA), a national assessment whose junior secondary mathematics tasks include geometry. At the broader policy level, improving how such content is taught relates to Sustainable Development Goal (SDG) 4 (Quality Education); in this article, SDG 4 is treated only as contextual motivation and not as an outcome examined by the study.

Digital technology offers ways of integrating spatial visualisation, interaction and exploration into geometry teaching (Saat et al., 2024; Schoenherr et al., 2024). A meta-analysis of visualisation interventions reported positive effects on mathematics learning (Schoenherr et al., 2024), and a meta-analysis of digital learning media reported a positive contribution to mathematics learning outcomes (Azkia et al., 2023). Augmented-reality (AR) and virtual-reality (VR) tools have been examined as ways of supporting mathematics learning and spatial skills (Bertrand et al., 2024; Yang et al., 2025), including in geometry classrooms (Rohendi et al., 2025). A systematic review of VR and spatial abilities identifies interaction as an important design feature of such environments (Gittinger & Wiesche, 2024). Spatial skills are malleable through training (Uttal et al., 2013), and students have been argued to need more practice with spatial thinking, for example in geoscience education (McLaughlin & Bailey, 2023).

Despite this evidence, three gaps motivate the present study. First, systematic reviews of geometry education at secondary level (Weigand et al., 2025) and student-perspective research on the design of digital mathematics learning environments (Weinhandl et al., 2025) provide the background for designing such media. However, development studies of GeoGebra-based, AR-based and other geometry visualisation media have mostly evaluated problem solving, conceptual understanding or visualisation with topic-specific measures (Hidayati & Murtiyasa, 2024; Nindiasari et al., 2024). We located a few studies that deliberately trace the features of a geometry medium to the competencies of a national assessment framework such as TKA and then examine students' performance on tasks aligned with that framework. Second, the virtual-museum metaphor has been associated with learning achievement in a meta-synthesis (Mercan et al., 2023), but to our knowledge it has rarely been used to organise two- and three-dimensional geometry exhibits for junior secondary students. Third, comparisons of digital media with conventional instruction do not isolate the contribution of interactivity itself; given the role of interaction identified by Gittinger and Wiesche (2024), a sharper test compares interactive 3D exploration with another visual but static medium. A gap therefore remains for an exploration-oriented learning environment whose 2D–3D representations, net exploration and practice tasks are explicitly linked to TKA indicators and whose outcomes are examined against a comparison condition that is itself visual.

Problem Statement: Expected and Actual Conditions

The expected condition is that junior secondary students can visualise, represent and reason about plane figures and solids well enough to solve the reasoning-oriented geometry tasks that presented by TKA. Reviews of secondary geometry education, however, describe a continuing gap between how geometry is taught and the reasoning that assessments require, including a predominance of procedural approaches and limited use of interactive visual media (Weigand et al., 2025), while digital tools that support spatial visualisation are advocated as one way of narrowing it (Saat et al., 2024; Schoenherr et al., 2024).

The actual condition in the school studied here was consistent with this gap. In the needs analysis (Study One, 64 Grade IX students), 55 students (85.94%) reported limited learning media, 52 (81.25%) reported difficulty understanding two-dimensional geometry, 50 (78.13%) reported difficulty with TKA-oriented reasoning and 40 (62.50%) reported difficulty visualising three-dimensional objects. The mathematics teacher reported that geometry lessons were generally conducted with a whiteboard, and the principal identified low mathematics TKA scores as a concern for the school. Before the intervention, the mean score on the 20-item TKA-aligned test was about 52 on a 0–100 scale in both classes (52.97 and 51.72); that is, roughly half of the items were answered correctly. These data come from one school and illustrate the problem; they are not an estimate of its national prevalence.

The distance between the expected and the actual condition therefore concerns two things: the availability of suitable visual media for exploring two- and three-dimensional objects, and students' readiness for TKA-type geometry tasks. The research problem addressed here is how a geometry learning medium can be designed from needs identified in this way and aligned with TKA indicators, and whether students who use it improve in geometric ability, in comparison with instruction that uses another visual medium.

Theoretical Framework and Design Conjecture

The design of Ruang-Geo draws on three complementary strands of research. First, spatial skills are malleable and can be improved through training involving active engagement with spatial representations (Uttal et al., 2013), while interaction has been identified as an important design feature of digital environments intended to support spatial ability (Gittinger & Wiesche, 2024). Second, Duval's (2006) theory of semiotic representation registers emphasises that mathematical understanding involves the coordination and conversion of different representations of the same mathematical object. In geometry, coordination between a three-dimensional object, its two-dimensional representation or net, and associated symbolic representations provides an example of such representational coordination. Third, the cognitive theory of multimedia learning suggests that appropriately designed combinations of visual and verbal information can support learning while managing cognitive processing demands (Mayer, 2009).

Together, these perspectives yield the following design conjecture: interactive 3D representation → active spatial manipulation → coordination between representations (3D object ⇄ net ⇄ symbolic representation) → improved spatial reasoning → improved performance on TKA-aligned geometry tasks. This chain is a design conjecture rather than a demonstrated causal mechanism. The present study measured the intervention and students' test performance but did not directly measure the intermediate processes. In addition, TKA-aligned practice may improve test performance through greater familiarity with the assessment format rather than through improved geometric ability; the study design does not fully distinguish between these possibilities. Because only six of the 20 outcome items concern nets (see Table 3), the outcome test can reflect this chain only in part.

The contribution claimed for this study is therefore a design and practical contribution rather than a new learning theory. Its principal innovation is pedagogical: Ruang-Geo uses a virtual-museum metaphor to organise two- and three-dimensional objects as explorable exhibits, and each object page brings together the interactive 3D model, its net and the symbolic information (properties, formulae and sample problems) so that students can practise converting between representations. Alignment with TKA indicators is treated not as an innovation in itself but as the way in which the practice tasks and the outcome measure were specified, so that learning could be evaluated against a recognised assessment framework (Table 2). The study also builds on the research team's earlier technology-based media: a GeoGebra-assisted realistic mathematics approach for numeracy among mathematics education students (Rimbatmojo, 2024), a mobile-learning medium for real analysis (Rimbatmojo & Sudrajat, 2025), and an AR medium for solid shapes in a primary school (Fasha et al., 2025). Specifically, the GeoGebra-assisted approach was designed for university students' numeracy and was not tied to junior secondary geometry; the mobile-learning medium addressed an advanced university topic (real analysis) and therefore did not deal with geometric representations; and the AR medium for solid shapes was implemented at primary level as a community-service activity. To our knowledge, none was designed around TKA geometry indicators for junior secondary students or combined two- and three-dimensional exhibits, nets and TKA-aligned practice in one environment. The incremental contribution of the present study is therefore twofold: (i) a documented junior secondary geometry learning environment whose features are traced to needs identified in Study One and to selected TKA indicators, and (ii) preliminary evidence on how students' geometric ability changed after its use relative to a static visual comparison condition.

The study addressed three research questions: (1) What learning needs and difficulties in geometry were identified in the needs analysis, and which design requirements followed from them? (2) How was Ruang-Geo designed, and how was its quality and feasibility judged by content experts, media experts, and a practitioner? (3) How did students' geometric ability change from pretest to posttest, and did learning outcomes differ between students taught with Ruang-Geo and those taught with worksheets and PowerPoint? In this study, the quantitative phase examines both within-group pretest-to-posttest changes and between-group differences in learning outcomes. It does not test equivalence or non-inferiority. Because the comparison condition (worksheets and PowerPoint) is itself a static visual medium, the comparison is best understood as interactive digital versus static visual instruction rather than digital versus non-digital teaching.

METHOD

Research Design

This study used a development approach following the ADDIE model within a sequential mixed-methods logic (Creswell & Plano Clark, 2018): findings of an initial needs analysis were intended to inform the design of an intervention, which was then tested quantitatively. The first strand was a descriptive needs analysis combining a student questionnaire with interviews with one teacher and one principal, analysed without formal qualitative coding. It is therefore a limited qualitative strand and should not be read as a full exploratory sequential design built on systematic qualitative inquiry. Three stages were linked to the three research questions: (1) Study One, a descriptive needs analysis based on a questionnaire and interviews with students, the mathematics teacher and the school principal, to identify learning needs and difficulties and to derive design requirements; (2) product development and validation,

in which the Ruang-Geo prototype was designed from those requirements and evaluated by content experts, media experts and a practitioner; and (3) Study Two, a quasi-experiment with a nonequivalent control group design to examine changes in geometric ability and compare them across instructional conditions. The logic of the sequence was: needs analysis → design requirements → product development → validation → implementation → learning outcomes → between-group comparison

Study One: Needs Analysis

Study One involved 64 Grade IX students (Classes IX F and IX G), one Grade IX mathematics teacher, and one school principal at SMP Negeri 1 Paguyangan, Brebes, Central Java, selected through purposive sampling; these were the same students who later took part in Study Two. Data were collected in May 2026 using semi-structured interview guides for the teacher and the principal and a questionnaire (survey) for the students, covering the availability of learning media, difficulties with two-dimensional geometry, difficulties with TKA-type reasoning, difficulties in visualising three-dimensional objects, and motivation or interest. Each issue was reported as a separate item, so the percentages are not mutually exclusive. Survey responses were summarised descriptively as frequencies and percentages of the 64 respondents. The interviews were summarised descriptively by interview topic (availability of learning media, geometry learning difficulties and TKA preparation) and compared with the survey results to check whether the two sources pointed in the same direction (data triangulation). No formal coding or thematic analysis was performed. The results were translated into design requirements and mapped onto Ruang-Geo features (Table 1).

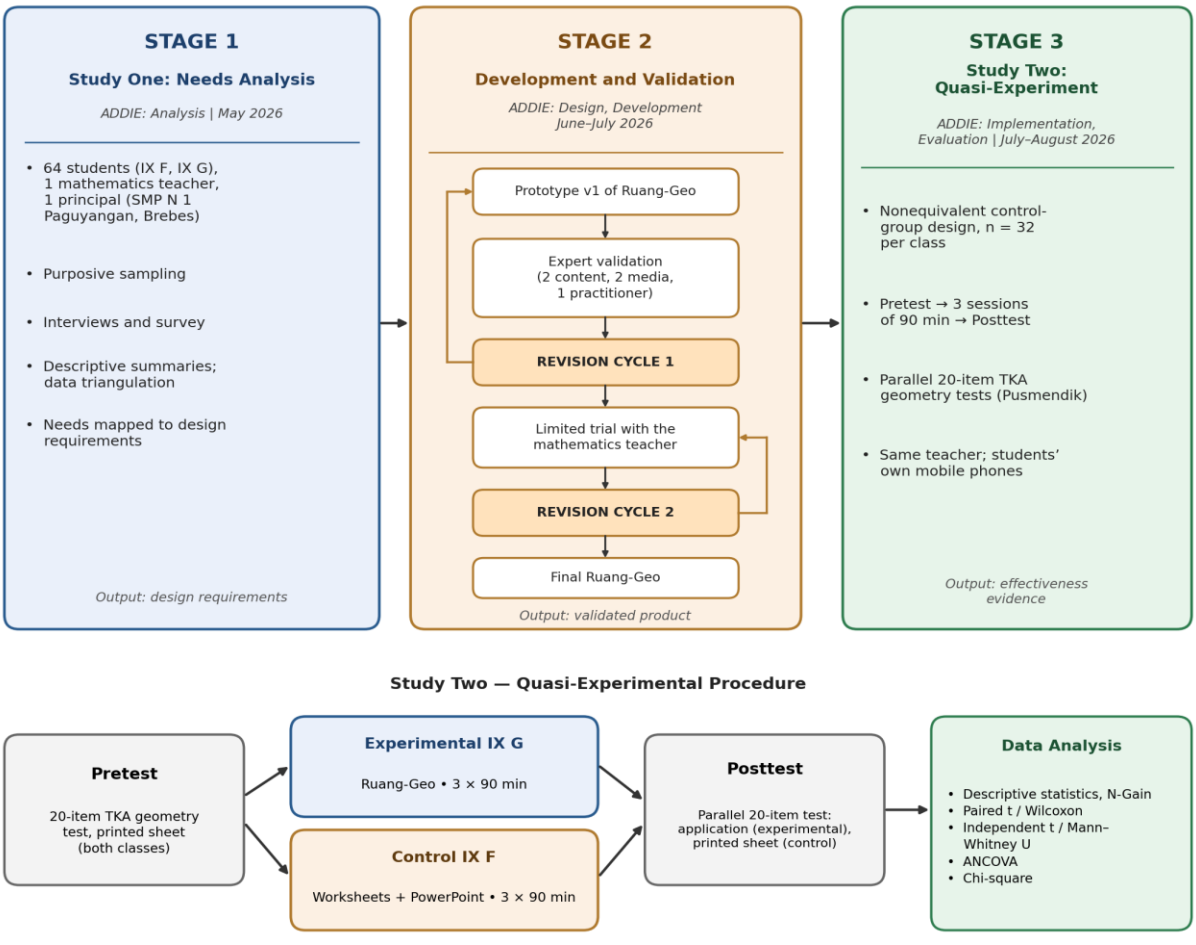


Figure 1. Research Flowchart of the Study, Showing the ADDIE Phases and Two Revision Cycles

Product Development Procedure

Development followed the ADDIE model (Analysis, Design, Development, Implementation, Evaluation). Study One corresponded to the analysis phase (May 2026). In the design and development phases (June–July 2026), a first prototype was built from the design requirements and improved through

two revision cycles (Figure 1). In Cycle 1, the prototype was reviewed by the validators and revised according to their comments. In Cycle 2, the revised prototype was tried out in a limited trial involving the mathematics teacher only (no students took part at this stage) and was revised again according to the teacher's feedback. The media-expert validation scores reported in this article were recorded after the revisions had been made. The implementation phase corresponded to Study Two (July–August 2026). Evaluation comprised both the formative stages (expert validation and the limited trial) and the summative effectiveness test. The main revisions made in each cycle are reported in the Results.

Participants

Study Two involved 64 students from two intact Grade IX classes at SMP Negeri 1 Paguyangan, Brebes, Central Java, Indonesia: 32 in the experimental class (Class IX G, taught with Ruang-Geo) and 32 in the control class (Class IX F, taught with worksheets and PowerPoint). The classes were assigned to the conditions based on availability rather than at random, and the geometry content the two classes had studied before the study was comparable. Because only one class was assigned to each condition, the treatment is confounded with class membership, a recognised threat to internal validity in nonequivalent control-group designs (Shadish et al., 2002; see Limitations). Product validation was conducted by five validators: two informatics lecturers (media/technical experts), two mathematics education lecturers (content experts) and one junior secondary mathematics teacher (practitioner/user).

Instruments and Procedure

Two instruments were used. The first was a 1–5 Likert-scale expert validation sheet assessing content appropriateness, alignment with TKA indicators, mathematical accuracy, language clarity, and systematic presentation (content experts); interface design, navigation, interactivity, 3D visualisation quality, and application stability (media experts); and practicality, appropriateness to student characteristics, usability, and usefulness (practitioner). The validation sheet itself was reviewed by the validators before it was used. Scores are interpreted as validity (content experts) and feasibility (media experts and practitioner) and represent expert judgement.

The second instrument was a geometry-ability test used as the primary outcome. In this study, geometric ability was operationally defined as the score on a test comprising items adapted from publicly available materials on the Pusmendik TKA SMP (junior secondary) mathematics portal (Pusat Asesmen Pendidikan [Pusmendik], n.d.) and aligned with the official TKA assessment framework. The test comprised 20 multiple-choice items, with parallel forms administered as the pretest and posttest. The item distribution covered the following indicators: relationships between angles and the Pythagorean theorem (2 items); congruence and similarity of plane figures (4 items); nets of three-dimensional figures (6 items); and perimeter and area of plane figures, including composite shapes (8 items). Because only the six net items concern three-dimensional representations directly, the test measures geometric ability on TKA-aligned tasks more broadly than spatial reasoning alone. Each correct answer scored one point, and the total was converted to a 0–100 scale (5 points per item). The pretest was administered on printed sheets to both classes, and the students were not given the answer key. The posttest used the parallel form: the experimental class completed it in the Ruang-Geo application, whereas the control class completed it on a printed sheet, so the mode of posttest administration differed between the classes (see Limitations). Because the items were taken from the official Pusmendik source and followed its blueprint, content validity rested on this official alignment. No additional content-validation procedure, reliability coefficient (Cronbach's α or KR-20), or item difficulty and discrimination indices were computed for this sample (see Limitations).

Research Procedure

The overall research flow, from the descriptive needs-analysis stage through product development and expert validation to the quantitative effectiveness test, is summarised in Figure 1, which also shows the ADDIE phases and the two revision cycles. In Study Two (July–August 2026), both classes studied plane geometric figures and curved-surface solids (cone, cylinder and sphere) for three sessions of 90 minutes each, taught by the same mathematics teacher with the same learning objectives and a comparable number of practice problems. In the experimental class, students used Ruang-Geo on their own mobile phones to explore the 2D and 3D galleries, object pages and quizzes; in the control class, the

teacher used PowerPoint presentations and printed worksheets covering the same content. The teacher aimed to keep practice time and support comparable in the two classes, but this was not verified by classroom observation. The pretest (printed sheets, for both classes) was administered before the instruction and the posttest (the application for the experimental class, a printed sheet for the control class) after it.

Data Analysis Technique

Expert validation scores were summarised descriptively as mean scores and interpreted with the following scale: 1.00–1.80 = very unfeasible; 1.81–2.60 = unfeasible; 2.61–3.40 = fairly feasible; 3.41–4.20 = feasible; 4.21–5.00 = very feasible (the corresponding labels 'valid' and 'very valid' were used for content experts). Because only two validators rated each aspect, individual ratings are reported alongside the means. Agreement with the maximum rating was also expressed with Aiken's (1985) V, $V = \Sigma(r - 1)/[n(c - 1)]$, where r is the rating, n the number of raters and c the number of scale categories; because of the very small panels, V is reported descriptively and no significance threshold is applied. Learning gains were expressed as the normalised gain, N-Gain = (posttest − pretest)/(100 − pretest) (Hake, 1998), and categorised as high (≥ 0.70), moderate (0.30–0.69) or low (< 0.30).

Baseline comparability was examined with an independent-samples t-test on pretest scores, reported with the 95% confidence interval of the difference and Hedges' g. Levene's test was used only to check homogeneity of variances and not as evidence of baseline equivalence. Shapiro–Wilk tests were used to assess the normality of the pretest, posttest, and N-Gain data. Improvement within each group was tested with a paired-samples t-test, confirmed with the Wilcoxon signed-rank test; the paired effect size is dz (mean difference divided by the standard deviation of the differences; Lakens, 2013), reported with the 95% confidence interval of the mean gain. No analysis plan was registered in advance. Because the classes were intact and not randomised, an ANCOVA of posttest scores with pretest as covariate, preceded by a test of the homogeneity of regression slopes (condition × pretest interaction), is the primary between-group analysis; because the slopes were not parallel, the common-slope estimate is reported together with the interaction test, and group differences at selected pretest values are reported as an exploratory analysis. Independent-samples t-tests of posttest scores and N-Gain (confirmed with the Mann–Whitney U test and, because the group variances may differ, with Welch's t-test) are reported first, for transparency, as unadjusted comparisons. Between-group effect sizes are Cohen's d (pooled standard deviation) and Hedges' g, interpreted descriptively with Cohen's (1988) benchmarks. The distribution of N-Gain categories was compared with Pearson's χ^2 test with Yates' continuity correction and checked with Fisher's exact test. The significance level was .05 (two-tailed). No equivalence or non-inferiority margin was specified in advance, so non-significant differences are not interpreted as evidence of equivalence (Lakens, 2017). A sensitivity power analysis (noncentral t distribution) was used to determine the smallest effect detectable with 32 students per group.

RESULTS

Needs Analysis Results (Study One)

The needs analysis involved 64 Grade IX students and was complemented by interviews with a mathematics teacher and the school principal. The student data indicated that limited learning media was the most frequently reported issue, identified by 55 students (85.94%). Difficulties in understanding two-dimensional geometry were reported by 52 students (81.25%), while 50 students (78.13%) reported difficulties with reasoning on TKA-oriented mathematics tasks. In addition, 40 students (62.50%) reported difficulties visualising three-dimensional geometric objects. A motivation or interest-related aspect was recorded for 25 students (39.06%).

The teacher interview highlighted the limited use of visual learning resources in geometry lessons, which were generally conducted using a whiteboard. According to the teacher, students needed learning media that could help them visualise geometric objects more concretely. The school principal also identified low mathematics TKA scores as a concern for the school and emphasised the need to prepare students for the assessment. These interview summaries provide contextual support for the student-

reported need for more accessible visual learning resources and greater preparation for TKA-oriented tasks.

Taken together, the findings informed the design requirements for Museum Digital Ruang-Geo, which were mapped onto its features as shown in Table 1. These design decisions were derived from the needs analysis; their contribution to students' geometric performance was examined separately in the subsequent evaluation study. Because Study One was descriptive and drew on one school, the needs it identified are specific to this context.

Table 1. Needs Identified in Study One, Design Requirements and Corresponding Ruang-Geo Features

Need identified	Evidence from Study One	Design requirement	Corresponding Ruang-Geo feature (Table 2)
Limited learning media	55 of 64 students (85.94%); teacher: geometry lessons were generally conducted with a whiteboard	A visual learning resource that students can use on their own mobile phones	Mobile application with Home, Materials, Quiz and Profile menus
Difficulty understanding two-dimensional geometry	52 students (81.25%)	Visual representations of two-dimensional shapes with support for formulae	2D Shapes Gallery with formula calculator (feature 3)
Difficulty with TKA-oriented reasoning	50 students (78.13%); principal: low TKA scores in mathematics	Practice tasks aligned with TKA indicators	Categorised Quiz, including a TKA category, and Contextual Practice Problems (features 5 and 6)
Difficulty visualising three-dimensional objects	40 students (62.50%); teacher: students need media that make objects more concrete	Interactive 3D representations connected to their nets	3D Shapes Gallery and Detailed Object Exploration with interactive 3D model and Net tab (features 2 and 4)
Motivation or interest	25 students (39.06%)	Elements that support motivation and orientation	Home menu with achievement points and geometry facts (feature 1)

Product Development Results

Museum Digital Ruang-Geo (hereafter Ruang-Geo) in Figure 2 was developed as a mobile application presenting an interactive geometry exploration gallery for junior secondary students. The application consists of four main menus (Home, Materials [a gallery of three-dimensional and two-dimensional shapes], Quiz, and Profile), with interactive 3D model visualisation as its key feature for every three-dimensional object. A description of the application's six main features and their relation to TKA indicators is presented in Table 2.

Table 2. Description of the Main Features of Ruang-Geo

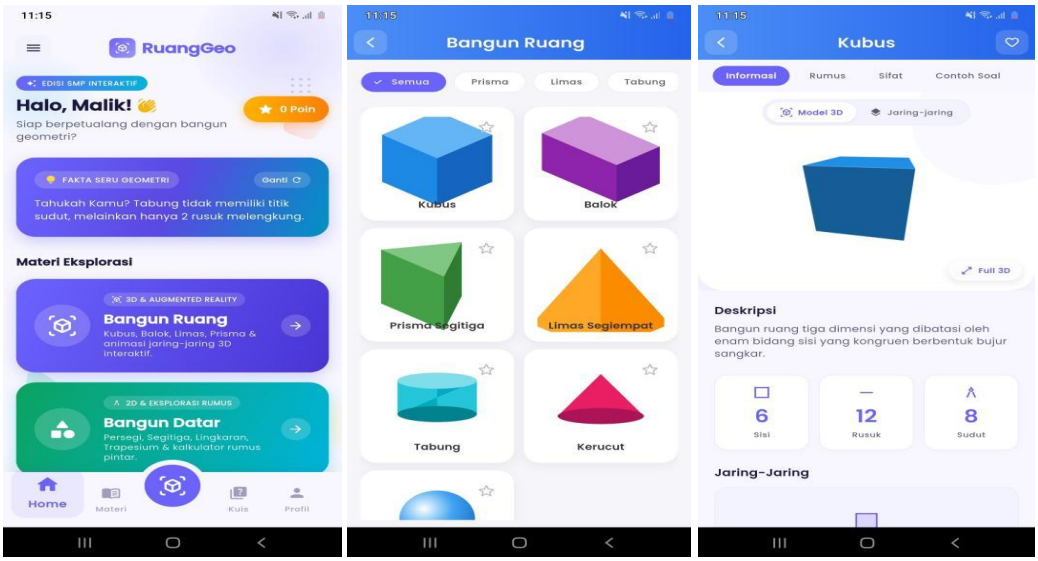
No	Feature/Menu	Function Description	Relation to TKA Indicators
1	Home	Personal greeting, student achievement points, interesting geometry facts, quick access to the Exploration Materials menu	Enhances students' initial motivation and orientation
2	3D Shapes Gallery	Collection of 7 three-dimensional shapes (cube, cuboid, triangular prism, square pyramid, cylinder, cone, sphere)	Representation and spatial visualisation of three-dimensional shapes
3	2D Shapes Gallery	Collection of two-dimensional shapes (square, triangle, circle, trapezoid, etc.) with a formula calculator	Representation of two-dimensional shapes and understanding of geometric formulas

No	Feature/Menu	Function Description	Relation to TKA Indicators
4	Detailed Object Exploration	Interactive 3D model of each three-dimensional shape with Information, Formula, Properties, Sample Problems, and Net tabs	Reasoning and spatial representation of three-dimensional shapes
5	Categorised Quiz	Question bank with 4 categories: Three-Dimensional Shapes, Two-Dimensional Shapes, Mixed, and TKA	Practice items aligned with TKA indicators
6	Contextual Practice Problems	Illustrated word problems with real-world contexts and a per-item timer	Contextual geometry problem solving consistent with TKA item characteristics

Table 3 relates the four indicators of the outcome test to the features in Table 2. Only the net items closely correspond to the interactive 3D features emphasised by the design conjecture; two indicators (6 of 20 items) have no dedicated exhibit. The instructional content in Study Two (plane figures and curved-surface solids) was also narrower than the domain covered by the test.

Table 3. Alignment between the Outcome-Test Indicators and Ruang-Geo Features

Test indicator	Items	Related Ruang-Geo feature (Table 2)	Remark
Nets of three-dimensional figures	6	3D Shapes Gallery and Detailed Object Exploration with the Net tab (features 2 and 4); Categorised Quiz (feature 5)	Closest match to the design conjecture
Perimeter and area of plane figures, including composite shapes	8	2D Shapes Gallery with formula calculator (feature 3); Categorised Quiz and Contextual Practice Problems (features 5 and 6)	Supported mainly through 2D exhibits and practice, not 3D manipulation
Congruence and similarity of plane figures	4	No dedicated exhibit	Practised, if at all, only through the Mixed and TKA quiz categories
Relationships between angles and the Pythagorean theorem	2	No dedicated exhibit	Practised, if at all, only through the Mixed and TKA quiz categories



(a) (b) (c)

Figure 2. (a) Home screen of Ruang-Geo; (b) 3D Shapes Gallery; (4) Detailed object exploration page (Cube) with 3D model, properties, and net

Expert Validation Results

Product validation used a 1–5 Likert-scale instrument administered to five validators. The results of the content expert, media expert, and practitioner validations are presented in Tables 4, 5, and 6.

Table 4. Content Expert Validation Results (2 Mathematics Education Lecturers)

No	Assessment Aspect	V1	V2	Mean	Category
1	Appropriateness of geometry content	4	5	4.5	Very Valid
2	Alignment with TKA indicators	4	4	4	Valid
3	Correctness of mathematical concepts	4	4	4	Valid
4	Language clarity	5	4	4.5	Very Valid
5	Systematic presentation of material	5	5	5	Very Valid
	Overall Mean	4.4	4.4	4.4	Very Valid

Table 5. Media Expert Validation Results (2 Informatics Lecturers)

No	Assessment Aspect	V1	V2	Mean	Category
1	Interface design/appearance	5	4	4.5	Very Feasible
2	Navigation and ease of use	4	4	4	Feasible
3	Application interactivity	4	5	4.5	Very Feasible
4	3D visualisation quality	4	5	4.5	Very Feasible
5	Application stability/performance	5	5	5	Very Feasible
	Overall Mean	4.4	4.6	4.5	Very Feasible

Table 6. Practitioner Validation Results (1 Junior Secondary Mathematics Teacher)

No	Assessment Aspect	Score	Category
1	Practicality of classroom use	4	Feasible
2	Appropriateness to junior secondary students' characteristics	5	Very Feasible
3	Ease of use (usability)	5	Very Feasible
4	Usefulness for geometry learning	5	Very Feasible
	Overall Mean	4.75	Very Feasible

Based on the validation results, Ruang-Geo was rated very valid by content experts (mean = 4.4), very feasible by media experts (mean = 4.5), and very feasible by the practitioner (mean = 4.75). Expressed as Aiken's V, the overall indices were 0.85 (content experts), 0.88 (media experts) and 0.94 (practitioner, one rater). The two validators in each expert panel differed by at most one scale point on any aspect; ratings were identical on three of the five content aspects and on two of the five media aspects. The lowest mean ratings (4.0) were given to alignment with TKA indicators and correctness of mathematical concepts (content experts), navigation and ease of use (media experts), and practicality of classroom use (practitioner). The media-expert scores in Table 5 were recorded after the revisions described below and therefore describe the revised prototype. Revisions were made in two cycles. In Cycle 1, the prototype was revised according to the validators' comments on content and technical aspects: the alignment of the content and practice items with the TKA indicators, the accuracy of mathematical concepts, and the navigation of the application. In Cycle 2, following the limited trial with the mathematics teacher, revisions were made to improve the clarity of the user interface and the practicality of classroom use based on the teacher's feedback. First, the interface was refined to provide a clearer visual distinction between the *2D Geometry* (plane figures) and *3D Geometry* (solid figures) menus, making it easier for students to identify and navigate to the relevant learning content. Second, an additional *Practice Questions* feature was introduced within the quiz function to allow students to complete independent practice beyond the core instructional materials before proceeding to the TKA practice section. This feature was intended to provide an additional opportunity for students to review and apply geometric concepts before attempting TKA-aligned tasks. These revisions focused on improving navigation, supporting independent learning, and ensuring a more coherent sequence from content exploration to practice and TKA preparation.

Because these results come from small panels (two experts per aspect and a single practitioner), they represent expert judgement and preliminary practitioner feedback. They indicate perceived validity and feasibility and do not demonstrate learning effectiveness or general classroom feasibility.

Product Effectiveness Test Results

The effectiveness test used a quasi-experimental nonequivalent control group design with 32 students in each of two intact classes. Pretest means were close (52.97 and 51.72); the difference of 1.25 points was small and not statistically significant (t(62) = 0.39, p = .698, 95% CI [−5.15, 7.65], Hedges' g = 0.10; computed from the values in Table 7). Levene's test indicated homogeneous pretest variances (F = 0.234, p = .630), which supports the use of pooled-variance tests but does not itself demonstrate baseline equivalence. Because each condition was a single intact class, differences between classes other than the instructional medium cannot be excluded. Descriptive statistics for the pretest, posttest and N-Gain of both groups are presented in Table 7. In addition, the posttest was completed in the application by the experimental class and on a printed sheet by the control class, so the effect of test mode cannot be separated from that of the instructional condition.

Table 7. Descriptive Statistics of Pretest, Posttest, and N-Gain

Class	Data	N	Mean	SD	Min	Max
Experimental (Ruang-Geo)	Pretest	32	52.97	12.17	30	75
	Posttest	32	70.16	10.04	50	85
	N-Gain	32	0.371	0.12	0.12	0.50
Control (Worksheet+PPT)	Pretest	32	51.72	13.42	30	75
	Posttest	32	68.13	7.70	55	85
	N-Gain	32	0.326	0.10	0.14	0.50

The maximum posttest score was 85 in both classes and the maximum N-Gain was 0.50, so no student reached the maximum possible score (100) or the high N-Gain category. The Shapiro-Wilk normality test indicated that most of the pretest-posttest data were normally distributed, except for the control class's posttest scores (p = .032) and the experimental class's N-Gain (p = .006), which were not normally distributed; the parametric test results were therefore confirmed with the corresponding non-parametric tests. The hypothesis-testing results are presented in Table 8.

Table 8. Hypothesis Test Results (Paired and Independent Samples Tests); effect sizes are d_z for paired tests and Cohen's d (pooled SD) for independent tests

Test	Group/Variable	t	Sig.	Effect size (d)
Paired	Experimental (Pre-Post)	13.992	<0.001	2.47
Paired	Control (Pre-Post)	10.988	<0.001	1.94
Independent	Posttest (Exp vs Control)	0.908	0.367	0.227
Independent	N-Gain (Exp vs Control)	1.626	0.109	0.407

Table 9. Supplementary Estimates (Confidence Intervals and Hedges' *g* Derived from the Summary Statistics in Table 7; ANCOVA Computed from the Raw Data)

Analysis	Estimate	95% CI	Test statistic	Effect size
Pretest difference (Ruang-Geo – control), points	1.25	[−5.15, 7.65]	t(62) = 0.39, p = .698	g = 0.10
Posttest difference, points	2.03	[−2.44, 6.50]	t(62) = 0.91, p = .367	g = 0.22
N-Gain difference	0.045	[−0.010, 0.100]	t(62) = 1.63, p = .109	g = 0.40
Mean pre–post gain, Ruang-Geo, points	17.19	[14.68, 19.69]	t(31) = 13.99, p < .001	d_z = 2.47 (g_z = 2.41)
Mean pre–post gain, control, points	16.41	[13.36, 19.45]	t(31) = 10.99, p < .001	d_z = 1.94 (g_z = 1.90)

Analysis	Estimate	95% CI	Test statistic	Effect size
Adjusted posttest difference (ANCOVA, raw data), points	1.33	[−1.36, 4.01]	$F(1, 61) = 0.98$, $p = .326$	partial $\eta^2 = .016$
Homogeneity of regression slopes (condition × pretest)	—	—	$F(1, 60) = 4.09$, $p = .048$	—

The paired-samples t-test showed significant improvement in both groups (Ruang-Geo: t(31) = 13.99, p < .001, mean gain = 17.19 points, 95% CI [14.68, 19.69]; control: t(31) = 10.99, p < .001, mean gain = 16.41 points, 95% CI [13.36, 19.45]; Tables 8 and 9). The paired effect sizes were very large by Cohen's (1988) benchmarks (dz = 2.47 and 1.94). These values reflect strong pre–post correlations (r ≈ .82 and .81, derived from the paired t statistics) and are not comparable to between-group effect sizes. The Wilcoxon signed-rank test was also significant in both groups (p < .001). Between-group comparisons with the independent-samples t-test showed no statistically significant difference in posttest scores (t(62) = 0.908, p = .367; mean difference = 2.03 points, 95% CI [−2.44, 6.50]; d = 0.227) or in N-Gain (t(62) = 1.626, p = .109; mean difference = 0.045, 95% CI [−0.010, 0.100]; d = 0.407), confirmed by the Mann–Whitney U test (posttest: p = .336; N-Gain: p = .091). Welch's tests, which do not assume equal variances (the posttest SDs were 10.04 and 7.70), led to the same conclusions (posttest: t(58.1) = 0.91, p = .368, 95% CI [−2.45, 6.51]; N-Gain: t(60.0) = 1.63, p = .108, 95% CI [−0.010, 0.100]). The confidence intervals include zero, so the data are compatible both with no difference and with a small-to-moderate difference between the conditions. The evidence therefore does not demonstrate that Ruang-Geo produced greater gains than the comparison condition.

Because the groups were not randomised and pretest and posttest scores were strongly correlated (r = .80), a baseline-adjusted comparison is the primary between-group analysis. An ANCOVA of posttest scores on condition, with pretest as covariate, was computed from the raw data. The adjusted posttest means were 69.81 (Ruang-Geo) and 68.48 (control); the adjusted difference was 1.33 points (95% CI [−1.36, 4.01]; $F(1, 61) = 0.98$, $p = .326$, partial $\eta^2 = .016$), consistent with the unadjusted results. However, the test of homogeneity of regression slopes was significant, $F(1, 60) = 4.09$, $p = .048$: the pretest–posttest slope was steeper in the Ruang-Geo class (0.68) than in the control class (0.47). The common-slope estimate therefore describes the group difference at only one point of the pretest distribution and should be interpreted with caution. In an exploratory follow-up of the model with the interaction term, the estimated difference was −1.36 points (95% CI [−5.10, 2.37]) one standard deviation below the mean pretest score (39.6), 1.32 points (95% CI [−1.30, 3.94]) at the mean (52.3) and 4.00 points (95% CI [0.28, 7.72], $p = .036$) one standard deviation above the mean (65.1). Any advantage of Ruang-Geo may therefore have been concentrated among students with higher pretest scores. This post hoc pattern rests on a borderline interaction, was not pre-specified, was not corrected for multiple comparisons and requires replication; it does not change the conclusion that the main between-group comparisons were not statistically significant. A sensitivity power analysis showed that with 32 students per group, α = .05 (two-tailed) and power of .80, the design could detect only effects of d ≥ 0.71. If the true effect were d = 0.41 (the N-Gain estimate), power would be about .36 and approximately 96 students per group would be needed for power of .80. The non-significant results are therefore compatible with a true small-to-moderate effect that this design could not reliably detect; the confidence intervals (for example, −2.44 to 6.50 points for the posttest difference) show the range of differences that remain plausible.

The distribution of N-Gain categories showed a higher proportion of students in the moderate category in the Ruang-Geo class (78.1%; 25 of 32 students) than in the control class (56.3%; 18 of 32 students), and a lower proportion in the low category (21.9%; 7 students vs 43.8%; 14 students); no student in either class reached the high category. This difference was not statistically significant (Pearson χ^2 with Yates' correction: $\chi^2(1) = 2.551$, p = .110; Fisher's exact test: p = .109; uncorrected $\chi^2(1) = 3.473$, p = .062). It is therefore a descriptive pattern and not evidence of statistically established superiority. The comparison of mean pretest–posttest scores and the N-Gain category distribution is presented in Figure 5.

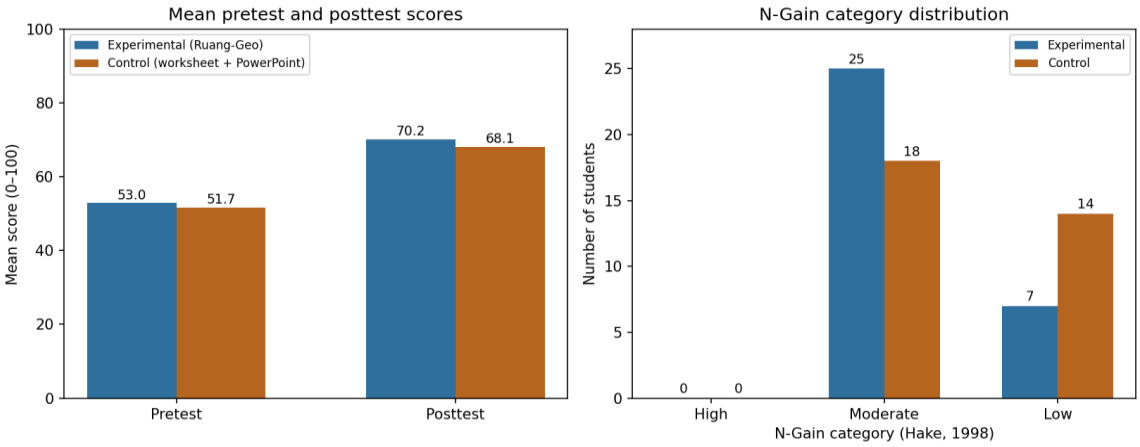


Figure 5. Mean Pretest-Posttest Scores and N-Gain Category Distribution

DISCUSSION

Students in both classes improved significantly in geometric ability from pretest to posttest, with large within-group effects (d_z = 2.47 and 1.94). This is consistent with evidence that visualisation-based and digital media can support mathematics learning (Azkia et al., 2023; Schoenherr et al., 2024), including AR studies of three-dimensional geometry (Rohendi et al., 2025). However, the design has no untreated comparison group, so the gains cannot be attributed to the media alone: regular instruction, maturation, familiarity with the TKA item format (although parallel forms were used) and regression to the mean may all have contributed. The gains show that both instructional conditions were associated with improvement and do not show that either medium caused it.

The difference favouring Ruang-Geo (N-Gain 0.371 vs 0.326; 78.1% vs 56.3% of students with moderate gains) was small. The 95% confidence intervals of the posttest and N-Gain differences include zero, and the posttest difference (2.03 points; g = 0.22) is small relative to the variability of scores. The direction of the difference is compatible with reviews suggesting that interactive, spatially exploratory environments may support spatial ability (Gittinger & Wiesche, 2024) and that virtual-museum activities may benefit achievement (Mercan et al., 2023). However, the first review concerns immersive virtual reality in engineering education rather than a mobile 3D viewer used by junior secondary students, so the transfer is tentative, and the present data neither confirm nor refute this proposition. Equivalence between the conditions was not tested, so the results do not show that Ruang-Geo is 'at least as effective' as conventional media either.

The baseline-adjusted analysis led to the same conclusion at the mean pretest score. However, the significant condition × pretest interaction suggests that the difference between conditions may vary with prior achievement, with an estimated advantage for Ruang-Geo only among students with relatively high pretest scores. This exploratory pattern should be tested in future studies with larger samples before any conclusion is drawn.

A possible pedagogical mechanism is embedded in the design of Ruang-Geo and is stated here as a hypothesis. In the worksheet-and-PowerPoint condition, students mainly receive static two-dimensional representations of three-dimensional objects, whereas in Ruang-Geo they rotate and manipulate interactive 3D models, move between the Information, Formula, Properties, Sample Problem and Net tabs, and apply what they explore to quiz items and contextual, timed word problems. The Net (jaring-jaring) feature, for instance, asks students to coordinate a 3D object with its unfolded 2D form, a conversion between representations (Duval, 2006) that relates to the net items of the TKA-aligned test (6 of 20 items; Table 3). Placing the model, its net and the symbolic information on one page was also intended to reduce the split attention that arises when related representations are separated (Mayer, 2009); cognitive load was not measured, so this too remains a design assumption.

This mechanism was not observed. The study collected no process-level data such as interaction logs, think-aloud protocols, or classroom observations, and the score data cannot distinguish an effect of

exploration and conversion between representations from an effect of greater exposure to TKA-style items. Future studies should collect such process data, for example using the Study One instruments or interaction logs, to test the design conjecture.

Several factors may explain why no statistically significant between-group difference was found, but none were tested directly. First, the comparison medium was itself visual-digital (PowerPoint), so the difference in learning experience was smaller than it would have been against purely non-digital instruction. Second, the study was underpowered for small-to-moderate effects: with 32 students per group, only effects of $d \geq 0.71$ could be detected with 80% power, whereas about 96 students per group would be needed to detect $d = 0.41$. Third, the data do not support a ceiling effect: the highest posttest score was 85 on a 100-point scale and the highest N-Gain was 0.50 in both classes, so all students had room for further gains, and a more plausible, though untested, explanation for the small between-group difference is the short intervention (three 90-minute sessions, 270 minutes in total) together with the other design limitations. Fourth, with one intact class per condition, class-level differences (teacher, peer environment, classroom conditions) are confounded with the instructional medium and may have added or masked differences. Fifth, the posttest was completed in different modes (in the application by the experimental class and on a printed sheet by the control class), so familiarity with on-screen testing may have favoured, or on-screen testing may have disadvantaged, the experimental class; the direction of any such bias is unknown. Sixth, the test and the product were only partly aligned (Table 3): angle relationships and the Pythagorean theorem, and congruence and similarity (6 of 20 items), had no dedicated exhibit in Ruang-Geo, so these items could hardly be sensitive to its distinctive features.

The results of Study One, the design of Ruang-Geo and the validation results form a coherent chain, but each link has a different evidential status. The needs reported by students were mapped onto features (Table 1), and the experts and the practitioner rated the resulting prototype favourably (Tables 4–6). The lowest mean ratings (4.0) concerned alignment with TKA indicators, the correctness of mathematical concepts, navigation and ease of use, and practicality of classroom use. These are the aspects on which the two revision cycles focused, and they are also the aspects most likely to determine whether teachers and students keep using an application outside a study. Because the ratings come from two experts per aspect and one teacher, and because Study One relied on student self-report and short interviews, the chain shows that the product is plausible and usable in this context; it does not show that the identified needs were met in terms of learning.

The very large paired effect sizes should also be read with care. The *dz* statistic divides the mean gain by the standard deviation of the individual gains, which is small when pretest and posttest scores are strongly correlated ($r \approx .80$); it is therefore not comparable with between-group effect sizes. In raw terms, both classes gained about 16–17 points on a 0–100 scale in three sessions. This is a meaningful change, but it was obtained with parallel test forms and without an untreated group. A reasonable reading is that both forms of instruction were accompanied by clear improvement and that any additional contribution of interactivity is small relative to the contribution of instruction itself, at least at this sample size and duration.

Ruang-Geo can therefore be described as a prototype judged valid and feasible by a small panel of experts and one practitioner, and as a promising learning medium. This study does not show that it yields greater learning gains than worksheets and PowerPoint, and equivalence has not been demonstrated; claims of equivalence or non-inferiority would require a design with a pre-specified margin (Lakens, 2017). SDG 4 provides the policy context for this work, but the study did not examine any outcome at the level of educational systems.

Limitations

This study has several limitations. (1) It was conducted in one school with one intact class per condition, so treatment is confounded with class and clustering cannot be modelled; the results should be read as preliminary and not as strong causal evidence. (2) Assignment was not randomised, and baseline comparability was assessed only descriptively and with a *t*-test on pretest scores. (3) The sample (32 per group) was small, the intervention was short (three 90-minute sessions), and no delayed posttest was administered. (4) No process-level evidence was collected to verify the proposed mechanism. (5) The validation panels were small (two experts per aspect and one practitioner), agreement was assessed

descriptively only, and the media-expert scores were recorded after revision, so pre-revision scores are not available. (6) The geometry test consisted of only 20 multiple-choice items taken from the Pusmendik TKA SMP portal and aligned with the official blueprint. No additional content validation, reliability coefficients, or item difficulty and discrimination indices were computed for this sample, and the equivalence of the parallel pretest and posttest forms was not empirically verified. Measurement error may therefore be considerable. In addition, the test sampled the broader TKA geometry domain rather than only the topics taught (plane figures and curved-surface solids), and only six items concerned nets of three-dimensional figures, so it may have been only partly sensitive to the instruction and to spatial reasoning in particular. (7) The posttest was completed in the Ruang-Geo application by the experimental class and on a printed sheet by the control class, whereas the pretest was printed for both; the effect of test mode could therefore not be separated from that of the instructional condition. (8) The assumption of homogeneous regression slopes in the ANCOVA was not met (p = .048), so the common-slope estimate should be interpreted with caution; the conditional analyses at selected pretest values are exploratory and were not corrected for multiple comparisons. (9) The comparison condition was a static visual medium and not a non-digital condition. (10) Study One was descriptive (survey frequencies and interview summaries without formal coding) and drew on one school, one mathematics teacher and one principal, so the needs it identified reflect this context and may not generalise. (11) Ruang-Geo has no dedicated exhibit for two of the four test indicators (6 of 20 items; Table 3), so the test covers more than the product was designed to teach.

Implications

The findings are preliminary, so the implications below are offered as considerations for design and practice rather than as evidence-based prescriptions.

1. Learning design.
   For learning design, the study shows that an object-centred organisation of geometry content, with one page per object combining the 3D model, its net and the symbolic information, is feasible to build, is judged acceptable by experts and can be derived transparently from a short needs analysis (Table 1). Designers who adopt this approach should specify before a study which test indicators each feature is meant to serve (Table 3), because the present product addressed nets and the measurement of plane figures more directly than congruence, similarity, angle relationships and the Pythagorean theorem. They should also test whether interactive features add anything beyond a well-designed static presentation, for example by comparing otherwise identical lessons with and without manipulable models.
2. Curriculum development.
   For curriculum development, the study illustrates how items from a national assessment blueprint can be used to define the geometry outcomes that a unit should address, so that nets, composite figures, congruence and similarity, and angle relationships appear explicitly in unit plans and not only in the end-of-year test. A set of digital exhibits such as Ruang-Geo is better treated as optional exploration and practice material within existing units on plane figures and solids than as a replacement for teacher-led instruction, because both conditions in this study were teacher-led and both improved. Unit plans should also allow for the difference between practising in an application and being tested on paper, since familiarity with the presentation mode may influence performance.
3. Educational policy.
   For educational policy, the results are too preliminary to justify recommendations about adoption at scale. They do bear on three practical points. First, schools preparing students for TKA need visual learning resources that students can reach on their own mobile phones; limited media was the most frequently reported need in this school. Second, school leaders and policy makers who encourage digital media for TKA preparation should ask for evidence from multi-class studies with valid and reliable measures, and not rely on expert ratings or pre–post gains alone. Third, support for teachers, such as shared item banks aligned with TKA blueprints and training in using interactive media, may

matter as much as the media themselves. SDG 4 provides the policy context for these points, but the study did not examine system-level outcomes.

Recommendations

1. For teachers.
   Teachers may use Ruang-Geo or similar media as a supplement for exploring three-dimensional objects and nets and for TKA-style practice, while keeping teacher explanation and feedback in place. Because the evidence does not show an advantage over well-prepared worksheets and slides, the choice of medium can be made on practicality, device availability, and students' interest. Teachers should pay attention to navigation and classroom practicality, the aspects rated lowest by the validators.
2. For developers.
   Developers should add exhibits for indicators that currently lack dedicated features (congruence and similarity; angle relationships and the Pythagorean theorem), improve navigation, and record interaction logs so that patterns of use can be linked to learning outcomes. A version of the application should also be tested that presents the same content statically, so that the contribution of interactivity can be isolated.
3. For researchers.
   Future studies should (a) use more classes or schools with random assignment and multilevel analysis; (b) extend the implementation period and add a delayed posttest; (c) use a fully reported and validated outcome instrument administered in the same mode to all students; (d) specify the analysis plan in advance, including a non-inferiority margin if equivalence is the aim; (e) collect process data such as interaction logs, think-aloud protocols and classroom observations to test the design conjecture; and (f) examine whether the effect depends on prior achievement, as the exploratory interaction in this study suggests.

CONCLUSION

This study produced Museum Digital Ruang-Geo, a virtual-gallery-based digital museum application that integrates interactive 3D visualisation, contextual practice problems, and categorised quizzes aligned with TKA indicators. The product was rated very valid by content experts (mean = 4.4) and very feasible by media experts (mean = 4.5) and the practitioner (mean = 4.75); these ratings reflect the judgement of small panels. Students in both the Ruang-Geo class and the worksheet-and-PowerPoint class improved significantly in geometric ability (d_z = 2.47 and 1.94), but the between-group differences in posttest scores and in N-Gain (0.371 vs 0.326; 95% CI of the difference [−0.010, 0.100]) were not statistically significant. The study therefore provides preliminary evidence that Ruang-Geo is valid and feasible and that students who used it improved in geometric ability. It does not provide statistically significant evidence that Ruang-Geo produces greater gains than worksheets and PowerPoint, and equivalence was not tested.

The contribution of the study is primarily a design and practical one: a documented, TKA-aligned virtual-museum learning medium for junior secondary geometry and a design conjecture (interactive 3D manipulation → conversion between representations → spatial reasoning → TKA-aligned performance) that future research can test. Implications for learning design, curriculum development and educational policy, and recommendations for teachers, developers and researchers, are given in the preceding sections.

ACKNOWLEDGMENTS

The authors would like to thank the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia for funding this research through the 2026 Beginner Lecturer Research (Penelitian Dosen Pemula, PDP) scheme, the Institute for Research and Community Service (LPPM) of Universitas Peradaban for administrative support, and the school, teachers, and students who participated in the field trial of Museum Digital Ruang-Geo.

DECLARATION OF GENERATIVE AI USE

The authors used Claude (Anthropic) to translate the manuscript from Indonesian into English and, during revision, to suggest wording for passages revised in response to reviewers' comments and to draft the response letters. The study design, data collection, data analysis, interpretation and conclusions are the authors' own, and no generative AI tool was used to collect, generate or analyse data or to generate references. The authors proofread and edited all AI-assisted text, verified its accuracy against their data and sources, and take full responsibility for the content of this article.

REFERENCES

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Azkia, N. F., Muin, A., & Dimyati, A. (2023). Pengaruh media pembelajaran digital terhadap hasil belajar matematika: Meta analisis [The effect of digital learning media on mathematics learning outcomes: A meta-analysis]. *JPMI (Jurnal Pembelajaran Matematika Inovatif), 6*(5), 1873–1886. https://doi.org/10.22460/jpmi.v6i5.18629

Bertrand, M. G., Sezer, H. B., & Namukasa, I. K. (2024). Exploring AR and VR tools in mathematics education through culturally responsive pedagogies. *Digital Experiences in Mathematics Education, 10*(3), 462–486. https://doi.org/10.1007/s40751-024-00152-x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Duval, R. (2006). A cognitive analysis of problems of comprehension in a learning of mathematics. *Educational Studies in Mathematics, 61*(1–2), 103–131. https://doi.org/10.1007/s10649-006-0400-z

Fasha, E. F., Rimbatmojo, S., Rizal, G., Aeni, K., & Millah, A. S. (2025). Implementasi media Augmented Reality dalam pembelajaran bangun ruang di Sekolah Dasar (SD) Ta'allumul Huda Bumiayu [Implementation of augmented reality media in learning solid geometry at Ta'allumul Huda Elementary School, Bumiayu]. *Empowerment: Jurnal Pengabdian Masyarakat, 8*(2), 140–150. https://doi.org/10.25134/empowerment.v8i02.11886

Gittinger, M., & Wiesche, M. (2024). Systematic review of spatial abilities and virtual reality: The role of interaction. *Journal of Engineering Education, 113*(2). https://doi.org/10.1002/jee.20568

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

Hidayati, A. N., & Murtiyasa, B. (2024). Efektivitas media pembelajaran GeoGebra untuk meningkatkan kemampuan pemecahan masalah matematis siswa pada materi transformasi geometri [The effectiveness of GeoGebra learning media in improving students' mathematical problem-solving ability on geometric transformation]. *JIPMat, 9*(2). https://doi.org/10.26877/jipmat.v9i2.481

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811678

McLaughlin, J. A., & Bailey, J. M. (2023). Students need more practice with spatial thinking in geoscience education: A systematic review of the literature. *Studies in Science Education, 59*(2), 147–204. https://doi.org/10.1080/03057267.2022.2029305

Mercan, G., Selçuk, Z. V., & Keskin, M. O. (2023). The impact of virtual museum on learning achievement: A meta-synthesis study. *Journal of Human and Social Sciences, 6*(2), 520–544.

Nindiasari, H., Pranata, M. F., Sukirwan, S., Sugiman, S., Fathurrohman, M., Ruhimat, A., & Yuhana, Y. (2024). The use of augmented reality to improve students' geometry concept problem-solving skills through the STEAM approach. *Infinity Journal, 13*(1), 119–138. https://doi.org/10.22460/infinity.v13i1.p119-138

Pusat Asesmen Pendidikan. (n.d.). *TKA Matematika SMP sederajat* [Mathematics TKA for junior secondary schools and equivalent]. Kementerian Pendidikan Dasar dan Menengah. https://pusmendik.kemendikdasmen.go.id/tka/tka/view/mata-pelajaran-wajib/smp

Rimbatmojo, S. (2024). Implementasi pembelajaran matematika realistik berbantuan Geogebra untuk meningkatkan kemampuan numerasi mahasiswa pendidikan matematika [Implementation of realistic mathematics learning assisted by GeoGebra to improve the numeracy of mathematics education students]. *Jurnal Jendela Pendidikan, 4*(4), 441–450. https://ejournal.jendelaedukasi.id/index.php/JJP/article/view/1058

Rimbatmojo, S., & Sudrajat. (2025). Pengembangan media pembelajaran analisis real berbasis mobile learning [Development of mobile learning-based real analysis learning media]. *Jurnal Jendela Pendidikan, 5*(3), 676–685. https://ejournal.jendelaedukasi.id/index.php/JJP/article/view/1755

Rohendi, D., Ramadhan, M. O., Abdul Rahim, S. S., & Zulnaidi, H. (2025). Enhancing student's interactivity and responses in learning geometry by using augmented reality. *Eurasia Journal of Mathematics, Science and Technology Education, 21*(1), em2559. https://doi.org/10.29333/ejmste/15796

Saat, N. A., Alias, A. F., & Saat, M. Z. (2024). Digital technology approach in mathematics education: A systematic review. *International Journal of Academic Research in Progressive Education and Development, 13*(4), 173–184. https://doi.org/10.6007/IJARPED/v13-i4/2295

Schoenherr, J., Strohmaier, A., & Schukajlow, S. (2024). Learning with visualizations helps: A meta-analysis of visualization interventions in mathematics education. *Educational Research Review*, 100639. https://doi.org/10.1016/j.edurev.2024.100639

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin, 139*(2), 352–402. https://doi.org/10.1037/a0028446

Weigand, H. G., Hollebrands, K., & Maschietto, M. (2025). Geometry education at secondary level – A systematic literature review. *ZDM – Mathematics Education, 57*(4), 829–843. https://doi.org/10.1007/s11858-025-01703-1

Weinhandl, R., Baldinger, S., & Kapplmüller, M. (2025). Educational design characteristics of digital mathematics learning environments from a student perspective. *International Journal of Mathematical Education in Science and Technology*, 1–27. https://doi.org/10.1080/0020739X.2025.2562275

Yang, Y., Du, W., & Mavrikis, M. (2025). Spatial skill development through augmented reality in mathematics education: A scoping review. *Digital Experiences in Mathematics Education*. https://doi.org/10.1007/s40751-025-00187-8