Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Augmented Reality, Physical Manipulatives, and Static Representations in Middle School Geometry: A Mixed-Methods Study of Spatial Skill Development

Yunus Öztürk
Bolu Abant İzzet Baysal University
Nuh Yavuzalp
nuhyavuzalp@gmail.com

Bolu Abant İzzet Baysal University

Research Article

Keywords: Augmented reality, Digital representations, Mental rotation, Physical manipulatives, Spatial visualization

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11224521/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Middle school geometry requires students to coordinate two-dimensional views with three-dimensional structures. Although augmented reality (AR) may support this work, comparisons with static materials alone do not show whether its potential value extends beyond access to a manipulable model. This mixed-methods study examined a marker-based AR application alongside instruction using static representations and instruction using physical unit cubes. Seventy-four seventh-grade students from two public middle schools completed five lesson hours on drawing views of unit-cube structures. The contrasts were conducted in different schools and analyzed as two parallel, two-group quasi-experiments rather than a single three-arm study. Spatial visualization and mental rotation were assessed before and after instruction; eight students from the AR groups were interviewed about their experiences. Both AR groups improved on both outcomes, while both comparison groups improved on spatial visualization. After adjustment for pretest scores, no between-group difference was statistically significant. The estimated effects favored AR in all four comparisons, but the study lacked the precision to determine whether AR outperformed either comparator. Interviewees described using rotation to inspect parts obscured in static drawings. These findings suggest that AR can support spatial learning during ordinary geometry lessons, but do not establish its superiority over physical manipulatives or static-representation instruction. The study foregrounds the fit between representational affordances and spatial tasks while identifying the need for larger, better-controlled comparisons.

1. Introduction

Spatial ability, that is, the capacity to generate, retain, transform, and interpret mental representations of objects and spatial relations, is among the most consistent cognitive predictors of achievement in science, technology, engineering, and mathematics (STEM) fields, and it continues to predict advanced STEM qualifications even after verbal and quantitative ability are taken into account (Lubinski, 2010; Wai et al., 2009). Crucially for instructional practice, the construct is malleable: meta-analytic evidence demonstrates that spatial skills respond to targeted training and that the resulting gains transfer to related tasks and persist over time (Uttal et al., 2013). The research question has accordingly shifted from whether spatial ability matters to how ordinary subject-matter teaching can be designed so that it develops spatial skills alongside curricular content.

Middle school geometry is a natural site for this agenda and also one of its persistent problem areas. A recurring curricular demand is that students relate a three-dimensional structure to the two-dimensional views it produces when seen from different directions, and conversely reconstruct a structure from the views given to them. This coordination requires learners to hold a configuration in working memory, reorient it, and reason about parts occluded from the current viewpoint (Kenneally & Bentley, 2024; Piri & Cagiltay, 2024). Because textbook diagrams and board drawings cannot be rotated, opened, or decomposed, the representational work that would make these relations visible is transferred to students' mental imagery, and the representation intended to support learning thereby becomes an additional source of difficulty.

Two instructional responses have developed largely in parallel. The longer-established is the use of physical manipulatives such as unit cubes, whose benefits for mathematics learning are meta-analytically supported but which cannot display dynamic change or interiors that are not directly accessible (Carbonneau et al., 2013; Supli & Yan, 2024). The more recent is augmented reality (AR), which registers virtual objects within the learner's real environment in real time and in three dimensions and is theorized to relieve the extraneous load of constructing that structure mentally (Azuma, 1997; Shaghaghian et al., 2024; Sweller et al., 2011).

These two lines of work have rarely been brought into direct contact. The great majority of classroom studies contrast an AR condition with curriculum-standard instruction based on textbooks and static representations, and such designs confound two distinct sources of influence: the AR condition supplies a manipulable three-dimensional referent and a novel digital technology, whereas the comparison condition supplies neither (Supli & Yan, 2024; Wang et al., 2025). Any advantage observed therefore cannot be attributed to the affordances of AR rather than to the mere availability of a model that can be turned, opened, and inspected, which an inexpensive set of unit cubes also provides. Schools allocating limited resources need to know whether AR delivers benefits beyond those obtainable with manipulatives costing a fraction as much, and claims about its cognitive mechanisms remain untestable without a pedagogically credible active control.

The present study addresses this gap by evaluating a researcher-developed marker-based AR mobile application against both comparison conditions within a single curricular topic: drawing the two-dimensional views of three-dimensional unit-cube structures from different directions. Because the two contrasts were implemented in different schools, they are analyzed and reported separately as two parallel two-group quasi-experiments, and the resulting confounding of school context with comparison condition is treated as a constraint on interpretation throughout. Outcomes are examined separately for spatial visualization and mental rotation, and interviews with students from the AR groups explore how they worked with the representations available to them. The aim is thus to examine under which task demands the affordances of AR are likely to matter, rather than whether AR is effective in general. The following research questions guided the study:

1. What change is observed in seventh-grade students' spatial visualization and mental rotation performance following instruction supported by the AR mobile application?
2. Once prior performance is controlled, is a difference detected between AR-supported instruction and instruction supported by static representations in students' spatial visualization and mental rotation?
3. Once prior performance is controlled, is a difference detected between AR-supported instruction and instruction supported by physical manipulatives in students' spatial visualization and mental rotation?
4. How do students describe their experience of the AR-supported learning process, and which features of the application do they identify as helping them work on spatial tasks?

2. Literature Review

2.1 Spatial Ability and Mathematics Education

Spatial ability is a multidimensional construct involving the formation, maintenance, transformation and interpretation of mental representations of objects and spatial relations (Uttal et al., 2013). For consistency, this manuscript uses spatial ability for the general capacity and spatial skills for the measurable sub-components examined here, spatial visualization and mental rotation. Two condition labels are likewise used throughout: static-representation instruction for curriculum-based teaching relying on static two-dimensional materials and a two-dimensional on-screen isometric drawing tool, and physical manipulatives for tangible three-dimensional materials such as unit cubes.

The two sub-skills examined here should not be treated as interchangeable labels for spatial ability as a whole, because they make partially distinct cognitive demands and may therefore respond differently to instructional media. Spatial visualization involves the analytic construction and decomposition of multi-step spatial structures, proceeding through systematic examination of parts, views and structural relations, whereas mental rotation involves the holistic transformation of an object across imagined orientations and places greater demand on continuous manipulation in working memory (McGee, 1979; Piri & Cagiltay, 2024). The distinction is consequential here, because the intervention tasks required students not merely to view three-dimensional forms but to coordinate front, side and top views and to infer cube units hidden from the viewing position; it also carries the differential argument developed in Section 2.3.

Malleability does not imply that all learners benefit equally from the same representational support. Middle school students frequently experience representational incoherence when required to translate static two-dimensional materials into three-dimensional mental models, particularly where the structure includes rear or interior components occluded from view (Koparan et al., 2023; Wu et al., 2024). The difficulty is further shaped by domain-specific prior knowledge, which is among the strongest predictors of learning outcomes (Simonsmeier et al., 2022) and which accordingly motivates the covariate-adjusted analytic strategy adopted here. The spatial difficulty addressed in this study is therefore not merely a technological problem of displaying three-dimensional objects; it is an instructional problem of helping learners coordinate multiple representations while reasoning about them.

From a mathematics education perspective, the task examined here is geometric as well as spatial. Drawing the views of a unit-cube structure requires students to reason about how a solid is composed, which cubes determine each outline, and which parts are hidden from a given direction; this reasoning is the target of the curriculum objective on geometric objects (Koparan et al., 2023; Wu et al., 2024). How digital representations help learners move between a three-dimensional object and its two-dimensional views is therefore a question about mathematical representation as well as spatial cognition, and the comparison of AR with static and physical representations in this study addresses this question. Within mathematics education, recent work has examined how the spatial affordances of AR and the feedback designed into AR environments shape students' mathematical inquiry (Riding et al., 2025).

2.2 Representational Support in Geometry: Physical Manipulatives and Augmented Reality

Physical manipulatives such as unit cubes have traditionally been used to address this difficulty. They provide haptic feedback supporting depth perception, require no devices, and their benefits for mathematics learning are supported by meta-analytic evidence, albeit under documented boundary conditions relating to instructional guidance and learner age (Carbonneau et al., 2013). They are nonetheless limited where a task requires demonstrating dynamic change, or where the structure of interest has an interior that is not directly accessible (Supli & Yan, 2024). Augmented reality offers a different set of affordances: it overlays virtual objects onto real-world environments in real time and in three dimensions (Azuma, 1997; Billinghurst & Duenser, 2012), scaffolding perspective taking and allowing learners to update mental models with immediate feedback (Shaghaghian et al., 2024). Systematic review evidence indicates that three-dimensional visualization tools confer an advantage over static methods in developing mental rotation (Piri & Cagiltay, 2024), and classroom studies in mathematics and geometry report gains in spatial skills and engagement (Bacca et al., 2014; Koparan et al., 2023; Velázquez & Méndez, 2021; Wu et al., 2024).

These affordances should not be read as an inherent advantage of AR over other media. Systematic reviews emphasize that AR is not always superior to established methods and that its effectiveness depends largely on the pedagogical and cognitive quality of the application design (Gonnermann-Müller & Krüger, 2025), while empirical findings indicate that AR may be effective for particular demands, such as dynamic graphic manipulation, without guaranteeing improvement across all dimensions of spatial reasoning (Wang et al., 2025). A recent scoping review of AR for spatial skill development in mathematics education likewise found that, although many studies reported positive effects, the findings were inconsistent, and it attributed these discrepancies to differences in tool design and intervention length and to limited attention to contextual factors, measurement approaches and learning processes (Yang et al., 2026). Where interfaces present excessive visual stimuli or misalign virtual objects with the physical world, they may increase rather than reduce cognitive demand through a split-attention effect (Ekleş & Doğru, 2024). The relevant question is therefore not whether AR is globally superior, but under which task demands its affordances become instructionally meaningful.

This brings the two lines of work into contact at the point where evidence is thinnest. The extent to which virtual manipulation can complement or approximate the tangible experience provided by concrete materials remains a critical empirical gap (Koparan et al., 2023; Supli & Yan, 2024), and the few studies comparing the two media directly are inconclusive. Yıldız and Tüzün (2011) compared a three-dimensional virtual environment with unit cubes and detected no significant difference in either spatial visualization or mental rotation, while Tosik Gün and Atasoy (2017), comparing AR combined with concrete materials against concrete materials alone, likewise found higher scores in the AR group without a significant between-group difference. Testing AR against an established physical comparator whose efficacy is independently documented (Carbonneau et al., 2013) is therefore necessary if claims about its added value are to be interpretable, and it is this requirement that motivates the active control condition employed here.

2.3 Theoretical Framework and Differential Expectations

The strands reviewed above converge into a single interpretive framework. Frameworks conceptualizing spatial ability (McGee, 1979; Uttal et al., 2013) define what is being developed: two measurable sub-skills, spatial visualization and mental rotation, which make partially distinct cognitive demands. Cognitive Load Theory (Sweller et al., 2011) and Multimedia Learning Theory (Mayer, 2014) define how instructional media are expected to influence this development, namely by offloading onto an external representation the extraneous demand of constructing and transforming three-dimensional structure mentally, thereby freeing working memory for the geometric reasoning that is the object of instruction. Because cognitive load was not measured in the present study, this account functions throughout as an interpretive lens rather than as a mechanism the study can test.

Applied to the two contrasts examined here, this framework yields differential expectations rather than a blanket prediction of AR superiority. Relative to static-representation instruction, both conditions address the same curriculum objective, but only AR permits continuous reorientation and inspection from arbitrary viewpoints; given the distinction drawn in Section 2.1, its load-relief advantage should therefore be most visible in mental rotation and least visible in spatial visualization, where step-by-step analysis of parts and views can be adequately supported by static representations. Relative to physical manipulatives, both media externalize the three-dimensional structure that would otherwise be held in mind, and because they are theorized to relieve broadly similar sources of load, no directional superiority is predicted; the empirical question is whether a difference can be detected at all. These are expectations to be examined rather than hypotheses the study is powered to confirm, and no claim of equivalence is advanced on the basis of a non-significant result.

3. Method

3.1 Research Design

The quantitative core is best characterized as two parallel two-group quasi-experiments with nonequivalent control groups and with pretest and posttest measurement, rather than as a single three-arm experiment. In School A the AR intervention was compared with static-representation instruction; in School B the same intervention was compared with physical manipulatives. Because the three conditions were not implemented within one school, school context and comparison condition are partially confounded, and the two comparisons are therefore analyzed and reported separately; a pooled four-group model is reported only as a sensitivity check. All lessons in all four classes were delivered by the same person (Section 3.3); this rules out differences between teachers as an explanation for between-group or between-school differences, although it introduces the risk of experimenter expectancy discussed in the Limitations. What remains at the school level is variation in student composition and institutional resources.

A qualitative strand was embedded within this core in accordance with the embedded mixed-methods experimental design (Creswell & Plano Clark, 2017). Structured observations ran concurrently with the

intervention, serving both fidelity monitoring and contextual documentation, whereas semi-structured interviews followed the posttests so that questions could probe the complete intervention. The qualitative data are positioned throughout as exploratory support rather than as independent confirmatory evidence (Figure 1).

Figure 1

Embedded Mixed-Methods Design and Sequence of Data Collection across the Two Schools

Note. SVT = Spatial Visualization Test (0–15); MRT = Mental Rotation Test (0–48). Each band shows one school, read left to right in temporal order; only the comparison condition differed between schools.

3.2 Participants

An initial pool of 99 seventh-grade students from two public middle schools in a central district of Türkiye took part, with 60 students in School A and 39 in School B. The schools were selected on practical grounds, including official approval, adequate technological infrastructure, access to at least two seventh-grade classes, and the willingness of the regular mathematics teacher to support the implementation; no socioeconomic data were collected, so the schools cannot be compared on this dimension. Because School B had a class set of unit cubes, it was designated for the comparison between AR and physical manipulatives, whereas School A was used for the comparison between AR and static-representation instruction. Using separate schools also reduced the risk of treatment diffusion, but it prevented the effect of the comparison condition from being separated from the school context. In both schools the doctoral researcher taught all lessons using a committee-approved lesson plan, with the regular mathematics teacher supervising alignment with the intended learning objectives (Section 3.3).

Two intact classes were included in each school, and assignment was made at class level but was not random. Both tests were first administered to all four classes; the class with the higher proportion of students owning a suitable mobile device was designated the experimental group, and students without a device were supplied with one so that none was excluded on grounds of access. In the analyzed sample, School A's AR class had a significantly higher mental rotation pretest score than the static-representation class (Section 4.1); closely similar starting levels across both outcomes therefore cannot be assumed. This departure from random assignment introduces a potential selection threat, since device ownership may covary with home resources or prior technology experience; it is revisited in the Limitations.

Only students completing both the pretest and the posttest entered the quantitative analyses, and no imputation was applied. In total 25 students were excluded for want of paired data, giving attrition of 25.0% in School A (AR 25.9%, static representations 24.2%) and 25.6% in School B (AR 29.4%, unit cubes 22.7%); exclusions resulted from absence on one or both test days. Attrition bias cannot be excluded,

particularly in School B, where the difference between groups was largest. The final sample comprised 74 students with a mean age of 13.2 years, balanced by gender overall (Table 1).

Table 1

Participant Flow and Gender Composition of the Analyzed Sample

School and group	Assigned	Excluded	*n* analyzed	Female	Male
School A, AR	27	7	20	8	12
School A, Static representations	33	8	25	13	12
School B, AR	17	5	12	5	7
School B, Unit cubes	22	5	17	11	6
Total	99	25	74	37	37

Note. AR = augmented reality. Assignment was at class level and was not random. Students were excluded where paired pretest and posttest data were unavailable.

For the qualitative strand, eight students were selected from the two AR groups after the posttests: in each school the mathematics teacher nominated two female and two male students who had participated regularly and could articulate their experiences, and those nominated took part only if they consented voluntarily. Achievement level was not a criterion. Interviews lasted 12 to 14 minutes and were conducted approximately one week after the posttests. Because no control-group students were interviewed, these interviews explore students' experiences with the application rather than supporting comparative claims.

3.3 Instructional Conditions and Implementation

The intervention addressed the objective of drawing the two-dimensional views of three-dimensional objects from different directions, within the Geometric Objects unit of the national Grade 7 curriculum. Implementation comprised five lesson hours, matching the curricular allocation, with approximately one hour of teacher-led instruction and four of practice across a single week. Pretests were administered three weeks before instruction in order to reduce pretest sensitization (Willson & Putnam, 1982) and posttests in the week following it; given this short dosage, the design deliberately measured proximal outcomes rather than distal transfer. The same content was covered in all groups, and every condition followed the same five-phase sequence: orientation to the assigned material, teacher-guided explanation of views, guided practice, independent practice coordinating the two representations, and post-intervention assessment. The opening lesson was delivered identically in all four classes, and the conditions diverged only in the representational medium used during orientation and practice.

All lessons in all four classes were taught by the doctoral researcher rather than by the regular classroom teachers, following the committee-approved lesson plan, with the regular mathematics teacher present in a supervisory capacity. This arrangement removes teacher characteristics as a source of between-group and between-school variation; however, because the researcher both developed the application and implemented every condition, experimenter expectancy is a potential source of bias (see Limitations). The application and instructional design had been refined through pilot testing beforehand (Özdemir & Yavuzalp, in review). Before practice began, AR students received procedural guidance on installing the application, scanning the marker and navigating among modes, while the control groups received parallel preparation for their materials.

In the AR groups, students examined a structure generated by the application, predicted the requested view, submitted their response and received immediate feedback, with the review function unlocking only after three consecutive incorrect answers, so that access to the AR view was withheld until the task had first been attempted unaided. In the static-representation condition, students worked on the same objective and task types using the textbook's comprehension activities and a commercial exercise book, completed on the interactive whiteboard with the isometric drawing tool published by the National Council of Teachers of Mathematics. This condition is therefore best described as curriculum-standard whole-class instruction supported by static representations and a two-dimensional on-screen tool rather than as entirely non-digital instruction, since the tool permits structures to be built on an isometric grid but not freely rotated. In the physical-manipulatives condition, students completed the same activities by constructing the structures from unit cubes, so that support came through haptic manipulation rather than virtual rotation or digital feedback.

Implementation was documented through a structured observation form (Appendix B) covering lesson timing, objective coverage, material functionality, student engagement, practice opportunities and technical issues. Across the observed lessons, the intended sequence, instructional time, content coverage and practice structure were implemented as planned. Because the form was completed in the AR classes only, and by the person who also delivered the lessons, it constitutes procedural documentation rather than an independently validated fidelity score.

3.4 The AR Application

The application is a researcher-developed, marker-based mobile AR tool built with the Unity engine and the Vuforia image-target library; its development, design rationale and pilot evaluation are reported in full in a companion paper (Özdemir & Yavuzalp, in review). Practice items are generated automatically rather than authored in advance, so that the supply of distinct unit-cube structures is effectively unlimited across twelve graduated difficulty levels; in addition to the standard practice screen, the application offers a timed mode (Figure 2b). For each structure the learner is asked in turn for its appearance from five directions, namely top, front, back, right and left, and answers by shading cells on a grid, with automatic evaluation and immediate feedback. The AR view is contingent on performance: after three consecutive incorrect responses a review button appears, and directing the camera at a printed marker

renders the structure on the page, where it can be inspected from any angle. Exposure to the AR view therefore varied with the errors each student made, and the AR condition is best described as automatically generated practice with immediate feedback in which AR inspection served as an error-contingent scaffold. The interface was deliberately minimal: one structure at a time, controls consolidated in a single region, decorative elements removed, and virtual objects anchored to the marker. These choices are consistent with cognitive load principles (Sweller et al., 2011) and with design recommendations later synthesized from empirical cognitive load studies of AR (Gonnermann-Müller & Krüger, 2025); they describe design intent rather than measured effects (Figure 2).

Figure 2

Interface of the AR Application

Note. The generated structure appears with the five queried directions labelled around it, and the learner responds by shading cells of the grid; the review button appears only after three consecutive incorrect responses. Interface text has been translated from Turkish.

3.5 Data Collection Tools

Two spatial ability measures were administered as pretests and posttests to all groups, and two qualitative tools were used with the AR groups only. The *Spatial Visualization Test* (SVT), developed by Winter et al. (1989) and adapted into Turkish by Yıldız and Tüzün (2011), presents unit-cube structures and requires students to relate an isometric depiction to its appearance from other directions, demanding analytic coordination of visible and non-visible relations. It comprises 15 items with five options each, scored one point per correct item with no partial credit, giving a range of 0 to 15.

The *Mental Rotation Test* (MRT), developed by Peters et al. (1995) as a redrawn version of the Vandenberg and Kuse test and likewise adapted by Yıldız and Tüzün (2011), measures the ability to rotate three-dimensional objects mentally. Each of its 24 items presents a target figure assembled from unit cubes with four alternatives, two of which are rotations of the target; students identified both, scoring one point per correctly identified alternative, giving a range of 0 to 48. This scoring credits partially correct items and is therefore more lenient than the rule of awarding a point only when both alternatives are correct (Peters et al., 1995). Because the target figures are themselves unit-cube assemblies, the measure is structurally close to the material used in all three conditions. Yıldız and Tüzün (2011) reported α = .97 for the SVT and α = .71 for the MRT; reliability was not estimated for the present sample, so these coefficients describe the instruments in their adaptation sample. Items are not reproduced here because the adaptation authors did not grant permission to do so.

The qualitative tools were the structured observation form described above, used as contextual rather than outcome evidence, and a 14-question semi-structured interview protocol (Appendix A) covering prior experience with AR, ease of use, perceived effects on solving spatial tasks, the role of AR in

visualizing hidden parts of objects, preferences for classroom and home use, motivational effects, and suggestions for other subjects.

3.6 Data Analysis

Quantitative data were analyzed using SPSS 26.0 and G*Power 3.1. Sample size was determined by the number of intact classes the schools could make available rather than by an a priori power analysis, so the precision actually achieved is quantified in the Results through a sensitivity analysis reporting the smallest effect each comparison could have detected. Post hoc power was not computed, because it is a monotonic transformation of the obtained *p* value (Hoenig & Heisey, 2001). Non-significant between-group results are interpreted as the absence of a detected difference rather than as evidence of equivalence.

Analyses proceeded in three stages: pretest scores were compared within each school as a descriptive baseline screen; within-group change was examined for every group; and, as the primary between-group evidence, analysis of covariance (ANCOVA) was conducted separately for each school and outcome, with the posttest as dependent variable, group as fixed factor and the corresponding pretest as covariate. ANCOVA was preferred because conditioning on pretest performance controls baseline differences directly rather than relying on the non-significance of a screen, and because prior knowledge strongly predicts learning outcomes (Simonsmeier et al., 2022).

Distributional assumptions were verified with the Shapiro–Wilk and Levene tests. For within-group comparisons, test selection was based on Shapiro–Wilk tests of the paired posttest-minus-pretest differences: paired-samples *t* tests were used when these did not depart significantly from normality, and Wilcoxon signed-rank tests otherwise. The School A mental rotation differences were non-normal in both groups; the other six paired difference distributions did not show a significant departure. For independent-group baseline screens, Mann–Whitney U was used for School A spatial visualization because the AR group's pretest scores departed from normality; independent-samples *t* tests were used for the other three comparisons, applying Welch's correction to School A mental rotation, where Levene's test indicated unequal variances, and using the equal-variance rows for the other two after non-significant Levene tests. For ANCOVA, homogeneity of regression slopes was tested by adding the group by covariate interaction, and residual normality with the Shapiro–Wilk test. Effect sizes were calculated for all comparisons, including non-significant ones, and are reported with adjusted mean differences and 95% confidence intervals (Cohen, 1988). Minimum detectable effect sizes were computed in G*Power for a fixed-effects ANCOVA with two groups and one covariate at $\alpha = .05$ and 80% power.

The qualitative data comprised the eight transcripts and the observation notes. Interviews were audio-recorded with student assent and written parental consent and transcribed verbatim; analysis was carried out on the Turkish transcripts, and only the excerpts reproduced here were translated and checked by a second researcher. An inductive content analysis approach was adopted (Miles et al., 2014). Two coders independently coded the corpus: the researcher who conducted the implementation and a thesis monitoring committee member with no involvement in the intervention. Intercoder

agreement reached 92%, and discrepant segments were revisited jointly until consensus was reached. Member checking was not conducted and saturation was neither sought nor assessed. The two strands were brought together only at the interpretation stage.

3.7 Ethical Statement

The study was conducted in line with the principles of the Declaration of Helsinki and was approved by the Bolu Abant Izzet Baysal University Human Research Ethics Committee (12 October 2023, No. 2023/446). Permission to conduct the research in schools was additionally obtained from the provincial directorate of national education and from the administration of each participating school. Informed assent was obtained from students and written consent from their parents; participation was voluntary, and students were told that they could withdraw at any stage without academic disadvantage and that their answers would not be graded. All personal information was anonymized and coded, and the de-identified data were stored securely. For educational equity, control-group students were given an AR-supported experience after the procedure was completed. To limit diffusion of treatment, the two contrasts were implemented in separate schools, and within each school the AR and comparison classes were taught in separate lessons and classrooms.

4. Results

4.1 Baseline Comparability and Within-Group Change

Baseline screening of the 74 analyzed students detected a difference in one of the four within-school comparisons. In School A, spatial visualization pretest means were nearly identical (AR: M = 4.85, SD = 1.69, n = 20; static representations: M = 4.84, SD = 2.25, n = 25); the Mann–Whitney test was not significant, U = 241.000, p = .835. Mental rotation pretest scores were higher in the AR class (AR: M = 26.35, SD = 5.47; static representations: M = 18.80, SD = 3.06), Welch's $t(28.30) = 5.520$, $p < .001$. In School B, neither baseline comparison was statistically significant: spatial visualization, AR M = 4.58 (SD = 2.88, n = 12) versus unit cubes M = 5.65 (SD = 1.94, n = 17), $t(27) = -1.194$, p = .243; mental rotation, 26.92 (SD = 4.32) versus 27.59 (SD = 6.31), $t(27) = -0.319$, p = .752. These baseline screens are descriptive: non-significance does not establish equivalence or remove selection bias. In particular, School A's mental rotation imbalance warrants caution; the corresponding pretest was included as a covariate in each primary between-group analysis.

All four groups improved in spatial visualization from pretest to posttest (School A: AR, $p < .001$, d_z = 0.96; static representations, $p < .001$, d_z = 1.03; School B: AR, $p < .001$, d_z = 1.54; unit cubes, p = .003, d_z = 0.83). For mental rotation, the AR groups improved in both schools (School A: Wilcoxon $z = -3.422$, p = .001, r = .77; School B: $t(11) = -4.612$, $p < .001$, d_z = 1.33), whereas change was not statistically significant in the static-representation group (Wilcoxon $z = -1.532$, p = .126, r = .31) or the unit-cube group ($t(16) = -1.561$, p = .138, d_z = 0.38). These within-group changes do not establish that one

instructional condition outperformed another; the pretest-adjusted between-group comparisons in Table 3 provide the primary evidence, and none was statistically significant.

Table 2

Descriptive Statistics and Within-Group Change from Pretest to Posttest by Group and Measure

Measure and group	*n*	Pretest *M* (*SD*)	Posttest *M* (*SD*)	Test statistic	*p*	Effect size
Spatial visualization						
School A, AR	20	4.85 (1.69)	8.20 (3.67)	$t(19) = -4.282$	< .001	$d_z = 0.96$
School A, Static representations	25	4.84 (2.25)	7.20 (3.14)	$t(24) = -5.155$	< .001	$d_z = 1.03$
School B, AR	12	4.58 (2.88)	7.17 (2.98)	$t(11) = -5.338$	< .001	$d_z = 1.54$
School B, Unit cubes	17	5.65 (1.94)	7.06 (2.61)	$t(16) = -3.429$	.003	$d_z = 0.83$
Mental rotation						
School A, AR	20	26.35 (5.47)	33.00 (7.37)	$z = -3.422$	.001	$r = .77$
School A, Static representations	25	18.80 (3.06)	22.96 (9.71)	$z = -1.532$	.126	$r = .31$
School B, AR	12	26.92 (4.32)	32.08 (6.45)	$t(11) = -4.612$	< .001	$d_z = 1.33$
School B, Unit cubes	17	27.59 (6.31)	29.94 (6.25)	$t(16) = -1.561$	.138	$d_z = 0.38$

Note. AR = augmented reality. Spatial visualization scores range from 0 to 15 and mental rotation scores from 0 to 48. Test selection was based on Shapiro–Wilk tests of paired posttest-minus-pretest differences: paired-samples *t* tests for $p > .05$ and Wilcoxon signed-rank tests for $p < .05$. For paired *t* tests, $d_z = |t|/\sqrt{n}$ with *n* the number of complete pairs; for Wilcoxon tests, $r = |Z|/\sqrt{n}$ with *n* the number of nonzero paired differences. Negative *t* and *z* values reflect the pretest–posttest order used in the paired test or SPSS's signed-rank convention, respectively; positive mean changes indicate improvement. Effect sizes are reported regardless of statistical significance.

4.2 AR versus Static-Representation Instruction

In School A, the AR class entered with higher mental rotation pretest scores than the static-representation class, Welch's $t(28.30) = 5.520$, $p < .001$ (Section 4.1). The pretest-adjusted comparisons

detected no statistically significant group effect on either outcome: ηp^2 = .030 for spatial visualization and ηp^2 = .049 for mental rotation (Table 3).

The adjusted analysis is treated as primary because it conditions on baseline performance, particularly important given the disparity in School A mental rotation pretest means. Pretest mental rotation predicted posttest performance, $F(1, 42) = 5.10$, $p = .029$. The adjusted group difference favored AR but was not statistically significant: 4.974 [95% CI: −1.818, 11.766]; ηp^2 = .049 (approximately d = 0.45). The group contrast remains confounded with intact class, and the School A mental rotation ANCOVA residuals departed from normality; this result therefore warrants caution. The within-group Wilcoxon results in Table 2 cannot resolve the between-group contrast.

Table 3

Covariate-Adjusted Between-Group Comparisons for Both Schools and Both Outcomes

Comparison and outcome	Adjusted *M*: AR vs comparison	Adjusted difference [95% CI]	*F*	*p*	ηp^2
School A: AR versus static-representation instruction					
Spatial visualization	8.195 vs 7.204	0.991 [−0.771, 2.753]	1.29	.263	.030
Mental rotation	30.186 vs 25.212	4.974 [−1.818, 11.766]	2.18	.147	.049
School B: AR versus physical manipulatives					
Spatial visualization	7.746 vs 6.650	1.097 [−0.265, 2.459]	2.74	.110	.095
Mental rotation	32.350 vs 29.753	2.597 [−1.420, 6.614]	1.77	.195	.064

Note. AR = augmented reality; CI = confidence interval. Adjusted means derive from analysis of covariance with the corresponding pretest as covariate; degrees of freedom were (1, 42) in School A and (1, 26) in School B. Positive differences favor the AR group.

4.3 AR versus Physical Manipulatives

In School B, the adjusted analyses detected no statistically significant between-group effect on either outcome (Table 3). Adjustment increased rather than decreased the estimated spatial visualization difference, because the AR group began with lower pretest scores, and the covariate was strongly related to the outcome in that model, ηp^2 = .631.

The adjusted spatial visualization effect in this comparison, ηp^2 = .095 or approximately d = 0.65, was the largest of the four estimates, although it was obtained from the smaller of the two samples; the corresponding mental rotation estimate was approximately d = 0.52. The confidence intervals for both outcomes include zero, so the results indicate neither a difference nor equivalence between the two conditions.

4.4 Assumption Checks and Statistical Precision

Homogeneity of regression slopes was satisfied in all four models (Table 4); the interaction in the School B mental rotation model approached conventional thresholds, $F(1, 25) = 3.11$, p = .090, and that result is accordingly interpreted with additional caution. Residual normality was satisfied for three models but violated for School A mental rotation, W = 0.90, p = .001. The ANCOVA result for this model is retained as the primary pretest-adjusted result and interpreted with caution in light of this violation and the small, intact-class comparison.

Table 4 also reports the sensitivity analysis quantifying what each comparison was equipped to detect. Fixing α = .05 and power at .80, School A could detect effects no smaller than d = 0.84 and School B no smaller than d = 1.05. All four adjusted effects, ranging from approximately d = 0.35 to d = 0.65, fell below the relevant threshold. Interpreted prospectively, a replication seeking 80% power for d = 0.50 would require approximately 128 students per comparison, and one targeting d = 0.40 approximately 199.

Table 4

Assumption Checks, Estimated Effects, and Minimum Detectable Effect Sizes

School (*N*) and outcome	Group × covariate interaction	Residual *W* (*p*)	ηp^2	Estimated *d*	Minimum detectable *d*
School A (45)					0.84
Spatial visualization	$F(1, 41) = 0.27$, p = .609	0.978 (.553)	.030	0.35	
Mental rotation	$F(1, 41) = 0.90$, p = .349	0.904 (.001)	.049	0.45	
School B (29)					1.05
Spatial visualization	$F(1, 25) = 0.30$, p = .591	0.971 (.586)	.095	0.65	
Mental rotation	$F(1, 25) = 3.11$, p = .090	0.976 (.728)	.064	0.52	

Note. Shapiro–Wilk statistics refer to the standardized residuals of each model. Estimated *d* values were derived from partial eta squared for the adjusted group effect. Minimum detectable effect sizes were computed in G*Power at α = .05 and 80% power given the sample available in each school, and therefore apply to both outcomes within a school.

A pooled analysis of covariance containing all four groups returned the same conclusion, with the group effect non-significant for both spatial visualization, $F(3, 69) = 1.69$, $p = .178$, and mental rotation, $F(3, 69) = 1.98$, $p = .125$; it is reported only as a sensitivity check.

4.5 Student Experiences of the AR-Supported Learning Process

Interviews with the eight purposively nominated students, together with the observation notes recorded in the AR classes, yielded three themes (Table 5). Because all interviewees came from the experimental condition and saturation was not assessed, these findings are exploratory accounts of the AR experience rather than comparative or confirmatory evidence.

Technical interaction and accessibility. All eight students encountered AR for the first time, yet found the interface intuitive and reported no technical difficulties, one noting that "*I didn't have any difficulty; I used it comfortably, since I'm already familiar with mobile applications*" (Std 1); the observation notes concur, with requests for help concentrated early and declining thereafter, suggesting that managing the technology took up little of the practice time after the first lessons.

Cognitive scaffolding and spatial reasoning. The most frequently expressed benefit concerned access to structure that static materials conceal. One student observed that "*being 3D, we cannot see the back of it on paper, but when we rotate it on that paper, we can see its back more carefully*" (Std 7), and another that "*sometimes there were parts at the back that could not be seen; when I saw them there, they came alive in my mind more*" (Std 3). Students also connected this support to the view-drawing task itself: "*with the front and back views, if I hadn't known, I would have looked at them the opposite way, but because the application was right there beside me, it helped me solve the questions more easily*" (Std 5), a reference to the reversal between opposite views that the task requires students to coordinate. Interpreted through the scaffolding account of Kenneally and Bentley (2024), such descriptions are consistent with dynamic visual support facilitating the construction of three-dimensional models, and students' examination of structures from multiple angles aligns with the graphics movement dimension operationalized by Wang et al. (2025).

Affective engagement and pedagogical acceptance. Some students described immediate feedback as turning concern about making errors into satisfaction at correct answers, one student reporting that it "*made me excited that I might make a mistake, but it feels very good when I do it correctly*" (Std 7). Five students preferred to share the experience with peers, one explaining that "*in class I can see what my friends and I are doing together, and I can enter a competition; it's nicer with them*" (Std 8). Not all agreed,

however: another found that "*because there was noise in class, I had trouble concentrating; I could focus better at home*" (Std 2). All expressed a wish to use AR in other subjects, including science, social studies and technology and design. Given the short instructional period, the absence of delayed measures and the purposive selection of interviewees, these preferences may partly reflect a novelty effect (Arici et al., 2019).

Table 5

Themes and Code Frequencies from Student Interviews and Observation Notes

Theme	Codes (*f*)
Technical interaction and accessibility	First-time AR use, intuitive interface (8); technical fluency, absence of difficulty (8)
Cognitive scaffolding and spatial reasoning	Facilitating representational transition (8); mental model construction (6); conceptual clarity (5)
Affective engagement and pedagogical acceptance	Interdisciplinary transfer (8); demand for curriculum integration (7); preference for collective classroom use (5); epistemic curiosity (4); achievement satisfaction (3); social mediation (3)

Note. n = 8. AR = augmented reality. Frequencies describe this purposive sample and do not indicate thematic saturation.

5. Discussion

All four groups improved significantly in spatial visualization, whereas mental rotation improved significantly only in the two AR groups; within-group effect sizes were large for the AR groups on both outcomes and for both comparison groups on spatial visualization. Once pretest performance was controlled, however, none of the four between-group comparisons reached statistical significance. Two questions therefore structure the discussion: what the pattern of adjusted effect sizes suggests about where the affordances of AR matter, and what the precision of those estimates permits the study to conclude.

5.1 Differential Effects across Spatial Skills

In School A the adjusted group effect was smaller for spatial visualization (ηp^2 = .030) than for mental rotation (ηp^2 = .049), but neither difference was statistically significant. This directional pattern is compatible with the differential expectations set out in Section 2.3; it does not establish that AR had a different effect on the two skills. The ordering did not recur in School B, where the spatial visualization estimate (ηp^2 = .095) exceeded the mental rotation estimate (ηp^2 = .064), although Section 2.3 predicted no directional difference for that contrast; the School A pattern should therefore not be read as a stable skill-specific effect of AR. Mental rotation is theorized to impose high extraneous load when the required

transformation must be performed unaided from static materials (Kenneally & Bentley, 2024), and AR could externalize that transformation, allowing it to be experienced through manipulation rather than held entirely in working memory (Wu et al., 2024), which corresponds to the graphics movement advantage identified by Wang et al. (2025). Spatial visualization, by contrast, proceeds analytically and can be supported to a useful threshold even by well-structured static representations.

A feature of the comparison condition is relevant to this tentative interpretation. The static-representation group did not work from wholly static materials: its practice activities used an isometric drawing tool on the interactive whiteboard, which supports precisely the analytic, part-by-part construction on which spatial visualization depends, while not permitting free rotation. The smaller adjusted estimate for spatial visualization than for mental rotation is consistent with an account based on the fit between task and technology, but neither outcome provides evidence of a differential treatment effect, and the comparison condition may have supported spatial visualization more than its label suggests. This account interprets a directional pattern rather than demonstrating an effect, for the reasons set out in Section 5.3, and replication across settings is required before it can be generalized.

5.2 AR and Physical Manipulatives

In School B, students using a mobile application on ordinary mobile devices, including their own, reached posttest levels in the same range as students working with unit cubes, a material whose efficacy for mathematics learning is independently documented (Carbonneau et al., 2013), and no difference was detected on either outcome. This is of practical interest, because a curriculum-aligned application requiring no specialist equipment could serve schools where access to class sets of physical materials is limited or logistically difficult (Supli & Yan, 2024). The adjusted spatial visualization effect here was the largest estimated in the study and both comparisons numerically favored AR, but the confidence intervals are compatible with a small disadvantage, no difference, or a moderate-to-large advantage for AR; establishing comparability would require equivalence or non-inferiority testing against a pre-specified margin (Lakens, 2017). The pattern aligns with the two prior studies that placed the two media side by side and likewise found the technology-supported group numerically higher without reaching significance (Tosik Gün & Atasoy, 2017; Yıldız & Tüzün, 2011).

5.3 Interpreting Non-significant Results under Limited Precision

All four between-group comparisons were non-significant. The confidence intervals around the adjusted differences were wide, extending from small disadvantages to substantial advantages for AR, so non-significance cannot be read as evidence that the conditions did not differ. The estimated effects, approximately d = 0.35 to 0.65, were all below the minimum detectable effect for the corresponding comparison and cannot be taken as evidence of a benefit either. The results therefore do not determine

whether AR was more effective than, or comparable to, either comparison condition. The estimates and their confidence intervals are reported so that they can inform the planning of adequately powered replications and future meta-analyses of comparable AR interventions.

5.4 Mechanisms Suggested by Student Accounts

Students described in everyday terms what the literature conceptualizes as cognitive scaffolding: an external representation that carries part of the transformation the learner would otherwise have to perform mentally (Kenneally & Bentley, 2024; Wu et al., 2024). The benefit students reported most consistently, gaining access to hidden and rear structure through rotation, is relevant to mental rotation, for which the adjusted School A estimate favored AR without reaching statistical significance. Because neither strand independently established an effect and cognitive load was not measured, this correspondence is suggestive only. Students' reframing of concern about errors as satisfaction at correct responses may also have reduced affective demands on working memory (Wu et al., 2024); given the short intervention, however, their positive accounts may partly reflect a novelty effect (Arici et al., 2019).

5.5 Contribution and Implications for Practice

Returning to the gap identified in the introduction, this study contributes in three ways. First, whereas most prior work compares AR only with static-representation instruction, it places the same AR application alongside both a static-representation condition and a physical-manipulative condition, allowing the differential expectations developed in Section 2.3 to be examined across two theoretically distinct contrasts within a single intervention. Second, it reports covariate-adjusted effect-size estimates with confidence intervals and minimum detectable effects for both contrasts; these can inform the planning of future studies and meta-analyses and add classroom data from middle school students to research on the fit between task and technology (Wang et al., 2025). Third, it shows that a curriculum-aligned AR application built around automatically generated practice items can be implemented on ordinary mobile devices within the instructional time allocated by the curriculum. The interviewed students, all of whom were using AR for the first time, reported no technical difficulty, and the observation notes recorded fewer requests for help after the early lessons; as this evidence comes from a small nominated sample and researcher-completed records, it is preliminary.

For practice, the findings are more consistent with targeted use than with wholesale adoption. Teachers may reasonably use AR for content requiring dynamic three-dimensional transformation, such as rotating a solid or coordinating views taken from different directions, where its affordances are best matched to the demand of the task, while retaining static representations and physical manipulatives where these already suffice or where haptic engagement is valued in its own right. Two features of the present implementation may be transferable to other classrooms: the application ran on ordinary mobile devices, including students' own, and required no specialist equipment, and access to the rotatable

model was withheld until a task had first been attempted unaided, so that the technology supported rather than replaced the mental work it was intended to develop.

5.6 Limitations

Several limitations should be considered when interpreting these findings. The first set concerns the design. Each contrast was implemented in only one school, so school context and comparison condition are partially confounded, and unmeasured school-level factors such as student composition and institutional resources may have contributed to the results. Assignment was not random but was based on the proportion of students owning a suitable device, a characteristic that may be related to home resources and prior technology experience. Attrition was about one quarter in each school and differed somewhat between groups in School B, so attrition bias cannot be excluded. Students were nested within intact classes while the analyses were conducted at the individual level; because each condition comprised a single class, class and condition are fully confounded and clustering could not be modelled.

The remaining limitations concern implementation, measurement and inference. The researcher who developed the application also taught all lessons, which introduces a risk of experimenter expectancy, and the observation form was completed by the same researcher in the AR classes only, so no fidelity data are available for the comparison conditions. Exposure to the AR view depended on students' errors and was not quantified, so the contribution of AR inspection cannot be separated from that of automatically generated practice with immediate feedback. The samples could reliably detect only effects of approximately $d = 0.84$ or larger in School A and $d = 1.05$ or larger in School B (Section 4.4). Reliability was not estimated for the present sample, the MRT scoring rule credited partially correct items, residual normality was violated in the School A mental rotation model, and cognitive load was not measured, so statements about load remain interpretive. The intervention lasted five lesson hours and included no delayed posttest, so the durability of the gains and a possible novelty effect could not be examined. Finally, the qualitative data came from eight teacher-nominated students in the AR groups only, the implementing researcher was one of the two coders, and cell sizes did not allow gender differences to be examined.

6. Conclusion and Suggestions for Future Research

Five lesson hours devoted to the two-dimensional views of three-dimensional structures were accompanied by substantial gains among seventh-grade students in spatial visualization in all four groups and in mental rotation in both AR groups. Comparison groups also improved, however, and after controlling for prior performance the study did not establish that AR outperformed static-representation instruction or that it matched physical manipulatives. Because the design included no group without instruction, these gains cannot be fully separated from retest and maturation effects; the results are therefore consistent with evidence that spatial skills respond to focused instruction delivered within

ordinary curricular time (Uttal et al., 2013), while leaving open which representational medium drives the gain.

The main contribution of the study is a set of covariate-adjusted effect-size estimates whose pattern is theoretically interpretable but imprecise. In School A, the estimated advantage for AR over static-representation instruction was larger for mental rotation (ηp^2 = .049) than for spatial visualization (ηp^2 = .030), a directional pattern compatible with, but not evidence for, the proposed role of dynamic transformation. In School B, the spatial visualization estimate was the largest of the four (ηp^2 = .095) and exceeded the mental rotation estimate (ηp^2 = .064), so the School A ordering was not repeated, and the question of whether AR can substitute for physical manipulatives remains open. None of the four adjusted comparisons detected a statistically significant group effect, so these estimates should inform rather than settle future comparisons.

Three directions follow. First, replications should be planned prospectively around a smallest effect of interest rather than around the estimates reported here, since effects obtained from small samples tend to be overestimated. The sample sizes given in Section 4.4 assume individual-level allocation; because intact classes are the realistic unit of assignment, they need to be inflated by a design effect reflecting the intraclass correlation, which will increase the number of classes required. Second, within-school three-arm designs analyzed with multilevel models accounting for the nesting of students within classes would remove the confound that constrains the present study, and combining such a design with a longer instructional period and a delayed posttest would address durability and novelty together. Third, the impact of AR on cognitive processes should not be assessed through performance tests alone: real-time cognitive load and visual attention should be measured directly, using validated load scales and, where feasible, eye-tracking, so that the mechanism invoked throughout this literature is tested rather than assumed (Gonnermann-Müller & Krüger, 2025). Logging students' use of the AR view would also allow its contribution to be separated from that of practice and feedback.

Statements and Declarations

Competing interests

The authors have no relevant financial or non-financial interests to disclose. The augmented reality application examined in this study was developed by the first author for research purposes. The application is not commercially available and generates no financial benefit for either author.

Data availability

The de-identified quantitative data supporting the findings of this study are not publicly archived because consent for open data sharing was not obtained from the participants and their parents or legal guardians. The quantitative data may be available from the corresponding author upon reasonable request, subject to ethical and institutional approval. The full interview transcripts are not available for sharing because they contain contextual information that may increase the risk of participant

identification. The spatial ability instruments cannot be redistributed because permission to reproduce the items was not granted by the authors of the Turkish adaptations.

References

1. Arici, F., Yildirim, P., Caliklar, Ş., & Yilmaz, R. M. (2019). Research trends in the use of augmented reality in science education: Content and bibliometric mapping analysis. *Computers & Education, 142*, 103647. https://doi.org/10.1016/j.compedu.2019.103647
2. Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators & Virtual Environments, 6*(4), 355–385. https://doi.org/10.1162/pres.1997.6.4.355
3. Bacca, J. L., Baldiris, S. M., Fabregat, R., & Graf, S. (2014). Augmented reality trends in education: A systematic review of research and applications. *Journal of Educational Technology & Society, 17*(4), 133–149.
4. Billinghurst, M., & Duenser, A. (2012). Augmented reality in the classroom. *Computer, 45*(7), 56–63. https://doi.org/10.1109/MC.2012.111
5. Carbonneau, K. J., Marley, S. C., & Selig, J. P. (2013). A meta-analysis of the efficacy of teaching mathematics with concrete manipulatives. *Journal of Educational Psychology, 105*(2), 380–400. https://doi.org/10.1037/a0031084
6. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
7. Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). Sage.
8. Ekleş, Y., & Doğru, M. (2024). Augmented reality in science education: Academic achievement, anxiety and cognitive load levels of secondary school 7th grade students. *Journal of Science Mathematics Entrepreneurship and Technology Education, 7*(2), 91–110. https://izlik.org/JA78PF68BP
9. Gonnermann-Müller, J., & Krüger, J. M. (2025). Unlocking augmented reality learning design based on evidence from empirical cognitive load studies: A systematic literature review. *Journal of Computer Assisted Learning, 41*(1), e13095. https://doi.org/10.1111/jcal.13095
10. Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*(1), 19–24. https://doi.org/10.1198/000313001300339897
11. Kenneally, C. D., & Bentley, B. (2024). A cognitive load approach to molecular geometries: Augmented reality technology and visuospatial abilities in chemistry. *Education Sciences, 14*(9), 1036. https://doi.org/10.3390/educsci14091036
12. Koparan, T., Dinar, H., Koparan, E. T., & Haldan, Z. S. (2023). Integrating augmented reality into mathematics teaching and learning and examining its effectiveness. *Thinking Skills and Creativity, 47*, 101245. https://doi.org/10.1016/j.tsc.2023.101245

13. 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
14. Lubinski, D. (2010). Spatial ability and STEM: A sleeping giant for talent identification and development. *Personality and Individual Differences, 49*(4), 344–351. https://doi.org/10.1016/j.paid.2010.03.022
15. Mayer, R. E. (Ed.). (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge University Press.
16. McGee, M. G. (1979). Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin, 86*(5), 889–918. https://doi.org/10.1037/0033-2909.86.5.889
17. Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.
18. Özdemir, Y., & Yavuzalp, N. (in review). Principled containment: How cognitive theory constrains design decisions in augmented reality prototyping for geometry education.
19. Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test: Different versions and factors that affect performance. *Brain and Cognition, 28*(1), 39–58. https://doi.org/10.1006/brcg.1995.1032
20. Piri, Z., & Cagiltay, K. (2024). Can 3-dimensional visualization enhance mental rotation (MR) ability? A systematic review. *International Journal of Human–Computer Interaction, 40*(14), 3683–3698. https://doi.org/10.1080/10447318.2023.2196161
21. Riding, K., Yang, Y., Clark-Wilson, A., Mavrikis, M., Geraniou, E., & Crisan, C. (2025). Rethinking affordances and feedback in AR environments to foster richer mathematical inquiry: Lessons from Touch the Derivative. *Digital Experiences in Mathematics Education, 11*(3), 507–533. https://doi.org/10.1007/s40751-025-00179-8
22. Shaghaghian, Z., Burte, H., Song, D., & Yan, W. (2024). An augmented reality application and experiment for understanding and learning spatial transformation matrices. *Virtual Reality, 28*, 12. https://doi.org/10.1007/s10055-023-00904-x
23. Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., & Schneider, M. (2022). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist, 57*(1), 31–54. https://doi.org/10.1080/00461520.2021.1939700
24. Supli, A. A., & Yan, X. (2024). Exploring the effectiveness of augmented reality in enhancing spatial reasoning skills: A study on mental rotation, spatial orientation, and spatial visualization in primary school students. *Education and Information Technologies, 29*, 351–374. https://doi.org/10.1007/s10639-023-12255-w
25. Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer. https://doi.org/10.1007/978-1-4419-8126-4

26. Tosik Gün, E., & Atasoy, B. (2017). The effects of augmented reality on elementary school students' spatial ability and academic achievement. *Education and Science, 42*(191), 31–51. https://doi.org/10.15390/EB.2017.7140
27. 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
28. Velázquez, F. C., & Méndez, G. M. (2021). Application in augmented reality for learning mathematical functions: A study for the development of spatial intelligence in secondary education students. *Mathematics, 9*(4), 369. https://doi.org/10.3390/math9040369
29. Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101*(4), 817–835. https://doi.org/10.1037/a0016127
30. Wang, L., Zhang, Q., & Sun, D. (2025). Exploring the impact of an augmented reality-integrated mathematics curriculum on students' spatial skills in elementary school. *International Journal of Science and Mathematics Education, 23*, 387–414. https://doi.org/10.1007/s10763-024-10473-3
31. Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal, 19*(2), 249–258.
32. Winter, J. W., Lappan, G., Fitzgerald, W., & Shroyer, J. (1989). *Middle grades mathematics project: Spatial visualization*. Addison-Wesley.
33. Wu, J., Jiang, H., Long, L., & Zhang, X. (2024). Effects of AR mathematical picture books on primary school students' geometric thinking, cognitive load and flow experience. *Education and Information Technologies, 29*(18), 24627–24652. https://doi.org/10.1007/s10639-024-12768-y
34. Yang, Y., Du, W., Mavrikis, M., & Geraniou, E. (2026). Spatial skill development through augmented reality in mathematics education: A scoping review. *Digital Experiences in Mathematics Education, 12*(2), 277–310. https://doi.org/10.1007/s40751-025-00187-8
35. Yıldız, B., & Tüzün, H. (2011). Effects of using three-dimensional virtual environments and concrete manipulatives on spatial ability. *Hacettepe University Journal of Education, 41*, 498–508. https://izlik.org/JA73PR33HB

Figures

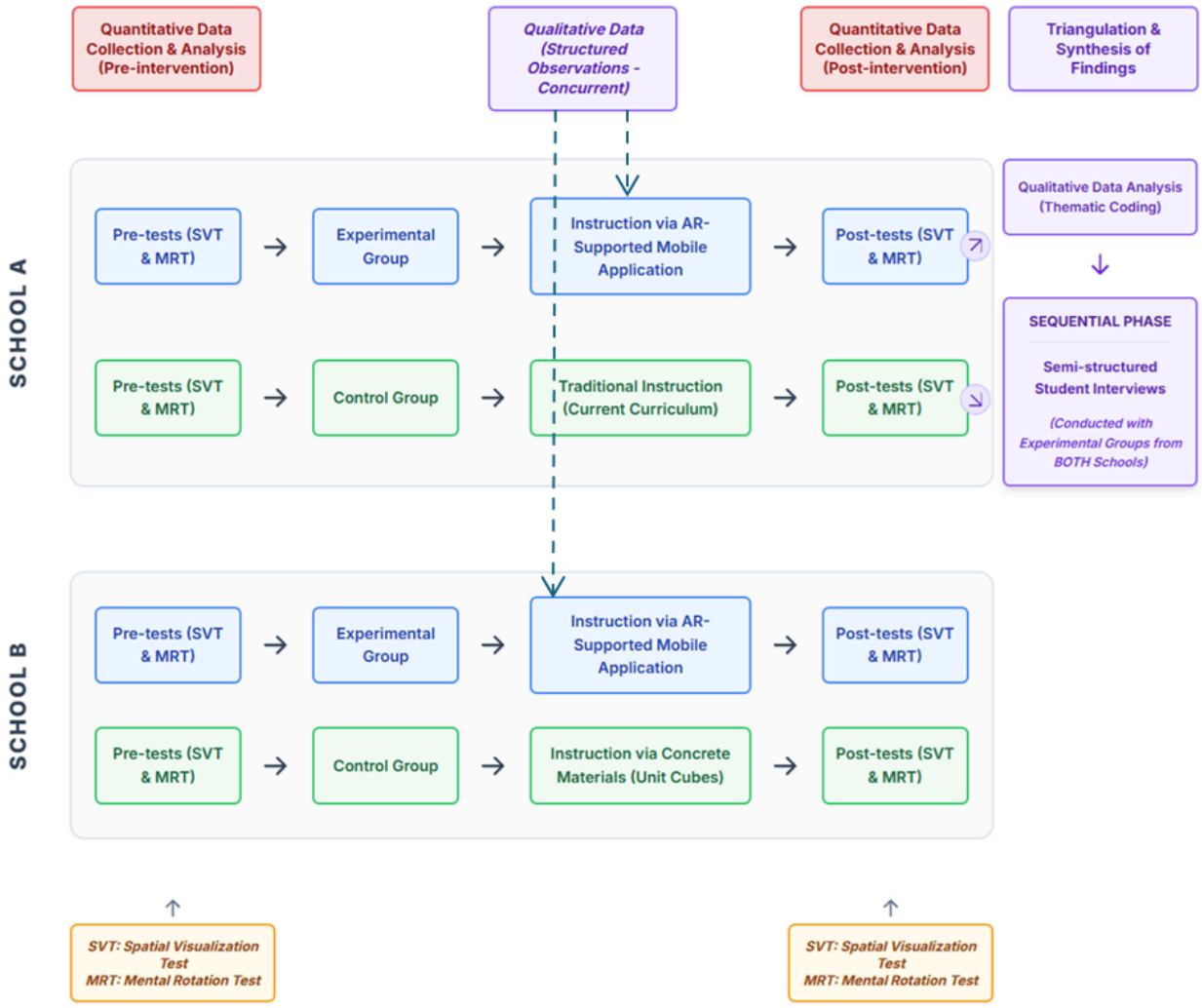
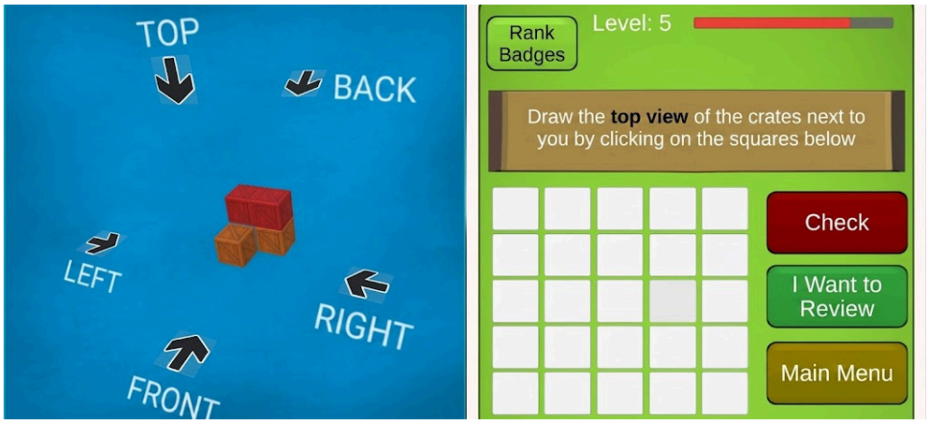


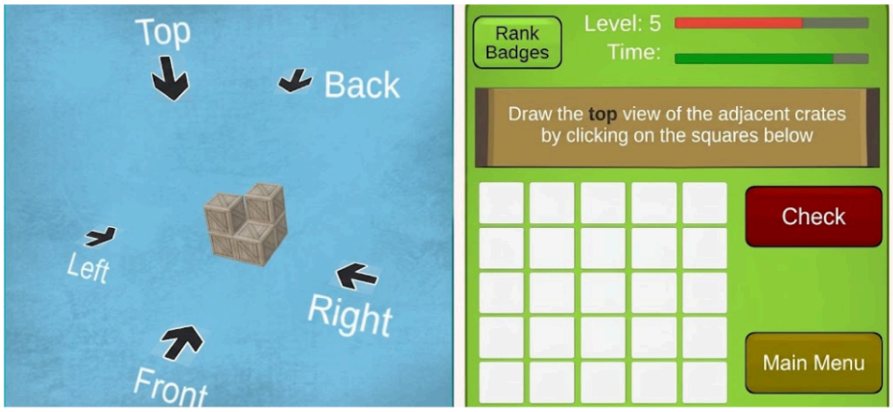
Figure 1

Embedded Mixed-Methods Design and Sequence of Data Collection across the Two Schools

Note. SVT = Spatial Visualization Test (0–15); MRT = Mental Rotation Test (0–48). Each band shows one school, read left to right in temporal order; only the comparison condition differed between schools.



(a) Standard practice screen



(b) Timed mode



(c) Augmented reality view

Figure 2

Interface of the AR Application

Note. The generated structure appears with the five queried directions labelled around it, and the learner responds by shading cells of the grid; the review button appears only after three consecutive incorrect responses. Interface text has been translated from Turkish.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx