Online Learning in Educational Research
Volume 6, Issue 4, 19 - 36.
e_ISSN: 2798-2351
https://www.journal.foundae.com/index.php/oler/index
DOI: 10.58524/oler.v6i4.1321

MathVerse M6: Exploring Students' Mathematical Critical-Thinking Performance in AR–Metaverse-Supported Learning

Ardi Dwi Susandi* Universitas Terbuka, INDONESIA
Puguh Darmawan Universitas Negeri Malang, INDONESIA
Siti Fatimah Universitas Pendidikan Indonesia, INDONESIA
Thesa Kandaga Universitas Terbuka, INDONESIA
Imam Rofiki Universitas Negeri Malang, INDONESIA
Entit Puspita Universitas Nusa Putra, INDONESIA
Hasan Basri Universitas Madura, INDONESIA

Article Info

Article history:

Received: June 16, 2026
Revised: August 11, 2026
Accepted: November 4, 2026

Keywords:

augmented reality
critical thinking sub-skills
metaverse
M6 learning model
TPACK

Abstract

Mathematical critical thinking involves distinct reasoning processes that are rarely evaluated separately in immersive technology-supported mathematics. This study aims to develop and formatively assess MathVerse M6 and describe overall and indicator-level change and exploratory phase-linked associations. An educational design research approach included 16 mathematics teachers in preliminary needs analysis and 118 Grade 9 students in a small trial (n = 28) and a large one-group pre–post-test trial (n = 90) across four schools in Cirebon, Indonesia. Three experts rated five artefacts, and classroom implementation and student responses were observed. Mean expert ratings were 3.48–3.82/4. The large-scale mean mathematics score increased from 48 to 82; mastery (score ≥75) increased from 0/90 to 75/90 (83.3%). Inference (+38 percentage points) and analysis (+35) showed the largest descriptive indicator changes; self-regulation (+18) showed the smallest. Previously reported phase and artefact correlations used only six and five aggregate pairs and remain exploratory. MathVerse M6 was judged suitable for trial and showed promising within-group changes, but the absence of a control group, clustered classes, single-item indicators, and inconsistent effect-size reporting preclude causal conclusions.

To cite this article: Susandi, A. D., Darmawan, P., Fatimah, S., Kandaga, T., Rofiki, I., Puspita, E., & Basri, H. (2026). MathVerse M6: Exploring students' mathematical critical-thinking performance in AR–Metaverse-supported learning. *Online Learning in Educational Research, 6*(4), 18-35. https://doi.org/10.58524/oler.v6i4.1321

INTRODUCTION

Mathematical critical thinking involves interpreting information, analyzing relationships, evaluating arguments, drawing warranted inferences, explaining conclusions, and monitoring one's reasoning. These processes are particularly important in geometry, where learners must connect symbolic expressions, diagrams, spatial configurations, and quantitative arguments. Recent work on mathematical critical thinking underscores the importance of inquiry, reflective reasoning, and explicitly structured learning opportunities (Ku et al., 2024; Mannan et al., 2026; Umam & Susandi, 2022; Vegas & Djukri, 2021; Wang & Abdullah, 2024). Treating critical thinking as a single undifferentiated outcome can obscure variation across these operations. Consequently, investigating sub-skill profiles may provide more diagnostically useful evidence than reporting a total test score alone. The six-part conceptual vocabulary is rooted in the expert-consensus account of critical thinking (Facione, 1990).

* **Corresponding Author**
Ardi Dwi Susandi, Universitas Terbuka, INDONESIA ✉ ardi.official11@ecampus.ut.ac.id

Augmented reality (AR) can overlay interactive three-dimensional representations on physical surroundings and allow students to manipulate geometric objects that would otherwise remain abstract. Reviews of mathematics applications report promise for spatial reasoning, visualization, engagement, and achievement, but also identify limitations in outcome measurement, intervention design, and access (Bulut & Borromeo Ferri, 2023; Erşen & Alp, 2024; Hidajat, 2024; Hidayat et al., 2026; İslim et al., 2024; Risdianto et al., 2022). Classroom studies have demonstrated that the dimensionality of the task, learning activity, and collaborative format may shape the value of AR rather than technology alone determining performance (Koparan et al., 2023; Walkington et al., 2024; Walkington et al., 2025; Walkington et al., 2026). Such findings motivate the deliberate alignment of digital affordances with mathematical reasoning tasks. Related findings in mathematical immersion, secondary-school AR, and the role of multiple representations appear in Çakıroğlu et al. (2023), Canbaz and Yalçın (2024), Çeken and Taşkın (2025), and Gonnermann-Müller and Krüger (2025).

Broader research on immersive and metaverse-oriented education offers an important qualification: visual immersion and novelty should not be equated with improved learning. Systematic reviews identify pedagogical integration, cognitive load, accessibility, teacher preparedness, and participant characteristics as recurrent design concerns (Alfaisal et al., 2023; De Felice et al., 2023; Feng et al., 2024; Maghaydah et al., 2024; Poupard et al., 2025; Pradana & Elisa, 2023; Xie et al., 2025). Meta-analytic evidence also suggests that educational effects vary across settings and research designs (Chang et al., 2022; Coban et al., 2022). In mathematics specifically, immersive environments can offer embodied opportunities for reasoning about perspective, scale, volume, and shape while requiring careful scaffolding (Bertrand et al., 2024; Wu et al., 2024; Zapata-Rodríguez et al., 2024). Additional syntheses document the affordances and limitations of educational immersion and the relationship between system design and learning contexts (Chamola et al., 2025; Gómez-Rodríguez, 2026; Hamilton et al., 2021; Leppink, 2024).

The M6 model was developed to structure mathematical critical thinking through six interconnected instructional phases (Susandi et al., 2022). MathVerse M6 extends this existing model by linking teacher-guided activities with AR and virtual collaboration affordances. The contribution proposed here is not the invention of an entirely new critical-thinking taxonomy; rather, it is the design and formative evaluation of a phase-aligned implementation and an exploratory account of how item-level outcomes correspond to theoretically targeted cognitive processes. Reviews of immersive education have called for greater attention to instructional design, learner interaction, and evidence quality (Camilleri, 2024; Miller et al., 2024; Sakr & Abdullah, 2024; Won et al., 2023). Prior implementations of M6 motivate the model-level framework but do not constitute an evaluation of the present AR-assisted application (Susandi et al., 2026).

In MathVerse M6, Focusing draws on prior knowledge and initial interpretation; Justifying engages learners in explaining and evaluating mathematical claims; Investigating supports hypothesis testing and inference; Presenting requires analysis and public justification; Evaluating invites reflection and monitoring; and Concluding consolidates explanations. These are instructional design intentions rather than experimentally isolated mechanisms. The six critical-thinking dimensions used in the assessment interpretation, analysis, evaluation, inference, explanation, and self-regulation were adapted from the framework described in the original M6 research materials. With only one test question per dimension, differences across sub-skills must be regarded as exploratory patterns rather than stable latent trait estimates (Istiqlal et al., 2026; Ku et al., 2024; Pereles et al., 2024).

Technological pedagogical content knowledge (TPACK) provides the overarching design rationale: technology features should serve subject-specific objectives through appropriate pedagogy. Contemporary reviews emphasize that TPACK is situated, dynamic, and responsive to the teaching context rather than a guarantee of learning gains (Brianza et al., 2024; Mishra et al., 2023; Schmid et al., 2024). In MathVerse M6, three-dimensional objects, virtual spaces, digital whiteboards, and teacher-controlled avatars are mapped to particular activities concerning the surface area and volume of prisms. The present study describes this integration, but does not estimate a formal TPACK measurement model or independently test its causal propositions.

Against this background, the study asks five descriptive and exploratory questions. RQ1: What evidence of product validity, implementation practicality, student responses, and within-

group change in mathematical critical thinking is observed after MathVerse M6 implementation? RQ2: What pre–post patterns appear across six single-item critical-thinking indicators, and how do they align descriptively with the M6 phase design? RQ3: How do implementation ratings vary across participating settings, and how do reported changes differ by prior-achievement category? RQ4: What exploratory association is observed between six aggregated phase-activity scores and the corresponding mapped indicator changes? RQ5: What exploratory association is observed between five product-validation ratings and the mapped indicator changes? Because there is no control condition, no randomized allocation, and no phase-isolation experiment, these questions concern implementation and associations, not causal phase effects or validated moderation.

METHOD

This study followed three stages of educational design research: preliminary investigation, prototype development, and formative assessment, using an M6 model informed by previous development work (Susandi et al., 2022). The evaluation combined teacher needs analysis, expert review of products and instruments, and two consecutive classroom trials. The effectiveness component adopted a pre-experimental one-group pre-test/post-test design. Consequently, observed changes can be described within the participating group but cannot be attributed uniquely to MathVerse M6. As a design framework, TPACK guided the relationship between technological resources, mathematics content, and pedagogical activities (Brianza et al., 2024; Mishra et al., 2023). The broad design stages follow the published educational design research framing (Nieveen & Folmer, 2013; Plomp, 2013).

Research Design

During preliminary investigation, teacher questionnaires and interviews addressed existing instruction, challenges in promoting critical thinking, and readiness for immersive resources. Prototype development produced five artefacts: an M6 model book, teacher guide, student workbook, AR–metaverse application, and mathematical critical-thinking test. Three experts then rated each artefact. A small-scale trial preceded a large-scale trial, with the manuscript indicating that revisions occurred between stages; however, a component-by-component change log was not available in the supplied record. Accordingly, this report does not specify unrecorded prototype iterations or unobserved design decisions.

Participants and Setting

The research was conducted in four junior high schools in Cirebon Regency, West Java, Indonesia. The preliminary needs analysis involved 16 mathematics teachers. Student trials involved 118 Grade 9 students in total: 28 students in one small-scale class and 90 students in three intact classes across three schools in the large-scale assessment. The sampling strategy was purposive rather than random. Consent and anonymization procedures were reported in the original research record. The reported number of participating students describes the two distinct trial samples; inferential pre–post results relate to the large-scale group of 90, not all 118 students. Because students were nested in intact classrooms and schools, individual-level test results may not be statistically independent. Neither an a priori power calculation nor classroom intraclass correlations were available in the supplied documentation.

Data Collection

Data sources comprised teacher questionnaires and semi-structured interviews for needs analysis; three experts' product and instrument ratings; model implementation observations (LPKM); phase-level student activity observations (LPAS); student response questionnaires (ARS); and a mathematical critical-thinking assessment administered before and after the large-scale implementation. The pre–post assessment covered the surface area and volume of rectangular and triangular prisms. The documentation does not establish interview coding procedures, intervention duration, allocation by gender, school-specific test scores, or the number of students within prior-achievement strata; these elements are therefore not reconstructed here.

Research Instruments

The expert-rating sheets addressed the format, content, language, visualization, accessibility, and related properties of the five products, depending on the artefact. LPKM captured how consistently the six M6 phases were enacted; LPAS documented observed student activity; ARS assessed student perceptions; and the mathematical critical-thinking test contained six questions, each linked to one indicator: interpretation, analysis, evaluation, inference, explanation, and self-regulation (Table 4). This single-item structure constrains sub-skill measurement precision: internal consistency cannot be estimated separately for each indicator, and item-level differences may reflect task difficulty or scoring variation. Although expert content ratings were available, no item difficulty, discrimination, rater agreement, or test–retest coefficients were supplied. The corresponding sub-skill findings are therefore presented as tentative descriptive signals rather than validated separate scales.

Data Analysis

Expert validation is reported as the arithmetic mean of ratings on the original four-point scale (Vr), not as Aiken's V: values in Table 5 are between 1 and 4 and cannot be generated by the normalized Aiken's V formula without rescaling. For this manuscript, Vr denotes the mean expert rating, with the original decision threshold of 3.00 retained. Implementation practicality (Pr), student activity (Ar), and responses were summarized using their documented four-point rubric criteria (Table 1). For the 90-student pre–post dataset, the original report supplied means, marginal standard deviations, a paired t statistic, and subgroup descriptions. Because paired difference scores were not available for reanalysis, the reported t statistic is presented as an unverified output rather than recomputed evidence. Normality of within-student differences, uncertainty intervals, and classroom-adjusted precision could not be independently established. Correlations involving only six phase pairs or five artefact pairs are retained, where stated, solely as illustrative ecological summaries, not confirmatory tests or individual-level predictions. Further analysis should account for the hierarchical sampling structure when class and school identifiers are available (Snijders & Bosker, 2012).

Table 1. Summary of Assessment Criteria Used in the Study

Assessment	Score	Criteria
Product validity	$3.0 \le Vr \le 4.0$	Valid
	$2.0 \le Vr < 3.0$	Moderately Valid
	$1.0 \le Vr < 2.0$	Invalid
Model implementation	$3.0 \le Pr \le 4.0$	High
	$2.0 \le Pr < 3.0$	Moderate
	$1.0 \le Pr < 2.0$	Low
Student activity	$3.5 \le Ar \le 4.0$	Active
	$2.5 \le Ar < 3.5$	Moderately Active
	$1.5 \le Ar < 2.5$	Insufficiently Active
Student response	$2.5 \le \bar{R}_i \le 4.0$	Positive
	$1.0 \le \bar{R}_i < 2.5$	Negative

RESULTS AND DISCUSSION

Preliminary Research Phase

Interviews identified a perceived need for clearer critical-thinking learning sequences, while the questionnaire documented variability in views of current materials and technology (Table 2). The percentages in the original summary were presented without underlying response counts for several items; they are consequently interpreted as approximate reported distributions, not recalculated frequencies from the 16 teachers. Importantly, 14 of 16 teachers expressed a strong perceived need for metaverse applications, whereas two raised restrictions associated with device policies or facilities. The teacher sample is small and purposive; these perceptions describe participating schools and do not provide population prevalence estimates.

The initial findings are broadly compatible with recent educational technology reviews: the availability of immersive representations does not by itself resolve instructional needs, and teacher preparedness and school infrastructure influence uptake (Maghaydah et al., 2024; Onu et al., 2024; Sakr & Abdullah, 2024; Xie et al., 2025). Consequently, MathVerse M6 was organized around teacher-guided problem exploration and reflection instead of relying solely on visualization. These data provide contextual design justification rather than independent evidence of effectiveness.

Table 2. Results of the Needs Analysis Questionnaire

Aspects	Indicators	Description of Results
Critical thinking and students' mathematical thinking in junior high school.	Students' Critical Thinking Skills	10% of teachers stated that students' critical thinking skills were in the "high" category, 26% in the "medium" category, and the remaining 64% in the "low" category.
	The potential of teaching materials in enhancing critical thinking skills	43% of teachers stated that the potential of the teaching materials used was "very unable" to improve critical thinking skills, 45% "unable," and 12% "able."
	The potential of a structured learning model in enhancing critical thinking skills	63% of teachers stated that there was a need for a learning model specifically used to improve students' critical thinking skills, while 37% stated that existing learning models were already able to accommodate this.
Teachers' needs for teaching materials based on the Metaverse.	The utilization of metaverse applications in learning	14 of 16 teachers (87.5%) stated that the utilization of metaverse applications was highly necessary, while 2 of 16 (12.5%) noted restrictions due to mobile phone policies or lack of facilities. 15 of 16 teachers (93.8%) had never used metaverse applications in teaching.
	The utilization of metaverse applications can enhance students' critical thinking skills	90% of teachers believed that metaverse applications could enhance students' critical thinking skills; 10% did not yet believe this.

Prototype Development

The prototype consisted of five artefacts that jointly operationalized the M6 sequence: the model book defined phase objectives; the teacher guide provided instructional orchestration; the student workbook supplied mathematics tasks; the application supported interactive and collaborative representations; and the test assessed outcomes. The application is described here as an AR-supported virtual learning environment based on the available feature descriptions; the record does not establish interoperability, persistence, or hardware specifications required to classify it as a fully realized metaverse platform.

Six phases structured the learning sequence. Focusing drew attention to prior geometric concepts and three-dimensional models; Justifying supported group discussion of mathematical rules; Investigating used simulated experiments and measurement to test ideas; Presenting asked students to communicate and defend proposed solutions; Evaluating involved correcting work and reflecting on reasoning; and Concluding consolidated concepts and applied them to further exercises. Phase descriptions specify intended opportunities for reasoning, rather than demonstrating that a single phase was responsible for a measured gain (Koparan et al., 2023; Walkington et al., 2024). Research on adaptive scaffolding suggests that the support learners need may differ according to available knowledge and representation demands (Kim et al., 2024).

The teacher and student books addressed the surface area and volume of rectangular and triangular prisms. The teacher book included implementation directions and suggested responses, whereas the student book presented exploration, discussion, problem-solving, and reflection tasks.

This differentiation is consistent with the need to align technological representations with a coherent progression of mathematical activity (Bulut & Borromeo Ferri, 2023; Wu et al., 2024).

The AR–metaverse application offered rotating three-dimensional objects, teacher-guided virtual interaction, simulation-based exploration, digital boards, and presentation and reflection spaces. The reported activities and student actions for each M6 phase are summarized in Table 3. The descriptions should be read as design specifications documented by the research team, rather than as independently observed evidence that all features were used with the same intensity or effectiveness.

Table 3. Metaverse Integration in the M6 Learning Model

Phase	Teacher Activities in the Metaverse	Student Activities in the Metaverse
Focusing on Initial Abilities	Explaining learning objectives through a video; recalling mathematical concepts using 3D models; using simulations to illustrate concepts of surface area of geometric solids.	Watching the teacher's video; exploring 3D mathematical objects; interacting to answer initial concepts.
Justifying Concepts	Inviting students to discover formulas; assigning exploration tasks; guiding discussions through avatars; assisting in concluding. Avatars in this context function as teacher-controlled representations facilitating real-time guidance and interaction within the virtual learning environment.	Engaging in discussion to discover formulas; exploring concepts through interactive simulations; discussing in groups; drawing conclusions.
Investigating Problems	Forming discussion groups; providing real-situation problems; guiding investigation with virtual measuring tools; providing guiding questions.	Forming groups; investigating problems through virtual experiments; recording observations; collaborating in solving problems and presenting results in 3D format.
Presenting Ideas	Facilitating presentation in a virtual hall; organizing interactive Q&A sessions; providing virtual whiteboards and digital presentation tools.	Presenting solutions using virtual whiteboards and 3D models; responding to other groups; documenting and sharing results.
Evaluating	Asking students to correct and reflect in a virtual reflection space; providing direct feedback through avatar; providing digital assessment rubrics.	Correcting work using simulation tools; explaining reasons for corrections; conducting individual reflection.
Concluding	Directing students to summarize concepts through digital boards; providing additional problems; creating digital summaries.	Formulating conclusions together; completing simulation-based exercises; recording and saving discussion notes.

The six-item assessment drew on the six dimensions of critical thinking used in the original M6 model (Table 4). Each dimension was represented by one question, making the sub-skill evidence task-specific. This design permits exploratory profiling but not reliable discrimination between latent sub-skills (Ku et al., 2024; Wang & Abdullah, 2024).

Table 4. Indicators and Descriptions of Critical Thinking Skills

Indicators	Description	Number of Questions
Interpretation	The ability to understand and express the meaning of various experiences, data, rules, procedures, or events.	1
Analysis	The ability to identify relationships between statements, questions, concepts, or other forms of expression.	1

Indicators	Description	Number of Questions
Evaluation	The ability to assess the credibility of information sources or the logical strength of an argument.	1
Inference	The ability to draw conclusions based on existing data, information, or evidence.	1
Explanation	The ability to express thinking results clearly and logically, including explaining the reasoning behind conclusions.	1
Self-Regulation	The ability to reflect on and re-evaluate one's own thinking process, including awareness of personal biases and weaknesses.	1
Total		6

Three experts evaluated each of the five artefacts using the original four-point criteria. Mean validation scores for the model book, teacher guide, student workbook, AR–metaverse application, and critical-thinking instrument are reproduced from the supplied manuscript in Table 5. Such ratings supply evidence of expert-perceived content and design quality; they are not a substitute for empirical psychometric validity or field effectiveness.

Table 5. Validity of the M6 Model Book Integrated with Metaverse (MathVerse M6).

Artefact	Aspect	Mean Rating	Overall Mean	Criteria
M6 model book	Format	3.58	3.59	Valid
	Content	3.60		Valid
	Language	3.58		Valid
Teacher book	Format	3.83	3.77	Valid
	Content	3.73		Valid
	Language	3.75		Valid
Student book	Format	3.83	3.82	Valid
	Content	3.87		Valid
	Language	3.75		Valid
AR–metaverse application	Visualization	3.25	3.48	Valid
	Ease of Access	3.58		Valid
	Content	3.60		Valid
	Animation	3.50		Valid
Critical-thinking instrument	Indicators-item alignment	3.33	3.56	Valid
	Language	3.67		Valid
	Appropriateness for Students' Thinking Level	3.67		Valid
	Content	3.58		Valid

The student outcome evaluation was conducted in two stages: a small-scale classroom trial involving 28 Grade 9 students and a subsequent large-scale trial involving 90 Grade 9 students from three intact classes. The evaluation summarized student scores, achievement relative to the minimum threshold of 75, implementation, student activity, and student responses. Table 6 presents the descriptive results for both trials, including the pre-test and post-test results available for the large-scale trial.

Table 6. Summary of Student Outcomes in Small- and Large-Scale Trials

Assessment Indicator	Small-Scale Trial	Large-Scale Trial:	
		Pretest	Posttest
Number of students (n)	28	90	90
Minimum score	64	20	66
Maximum score	100	70	100
Score range	36	50	34

Assessment Indicator	Small-Scale Trial	Large-Scale Trial:	
		Pretest	Posttest
Mean score	77	48	82
Standard deviation (SD)	Not reported	14.34	9.85
Students scoring ≥ 75, (n) (%)	21 (75.0%)	0 (0%)	75 (83.3%)
Implementation score (Pr)	3.3	—	3.5
Student activity score (Ar)	3.4	—	3.4
Student response score	3.6	—	3.7

Note. SD for the small-scale trial was not reported in the original data. Dashes (—) indicate that the corresponding values were not reported for that assessment stage.

Small-scale trial. Twenty-eight Grade 9 students participated in the preliminary classroom trial. The mean implementation score was Pr = 3.3 (high implementation), the mean observed student activity score was Ar = 3.4 (moderately active, see Table 1), and the student response score was 3.6 (positive). Twenty-one students (75.0%) scored at least 75, and the overall mean score was 77 (Table 6). Because the reported data do not include a corresponding small-scale pre-test, this result should not be interpreted as evidence of a pre–post gain.

Large-scale trial. The subsequent trial included 90 Grade 9 students from three intact classes. The implementation score was Pr = 3.5 (high implementation), the observed activity score was Ar = 3.4 (moderately active under the stated rubric), and the student response score was 3.7 (positive). Table 6 reproduces the documented marginal distributions. The mean score increased from 48 to 82, a change of 34 score points; 0 of 90 students met the threshold of 75 at pre-test, compared with 75 of 90 students (83.3%) at post-test. These are within-group differences; absent a control group, other explanations including concurrent instruction, maturation, familiarization, and repeated testing remain possible.

The original analysis reported a paired-samples t value of 18.35 (df = 89) and marginal Kolmogorov–Smirnov p values of .124 (pre-test) and .206 (post-test). The latter do not establish normality of paired differences, which is the relevant assumption for a paired t-test. In addition, the reported Cohen's d = 1.21 does not follow transparently from the stated pooled-SD formula: using the displayed marginal means and SDs would yield approximately 2.76, whereas t/sqrt(90) gives the distinct paired-difference effect-size metric dz ≈ 1.93. Rather than selecting a preferred value without the paired records, we omit a definitive standardized effect size. The original inferential result and uncertainty should be checked against the underlying data, including clustering by class. The observed 34-point mean change remains descriptively supported by Table 6.

For the six item-linked critical-thinking indicators, the supplied summary reported increases of 38 percentage points for inference (41% to 79%), 35 for analysis (43% to 78%), 31 for interpretation, 28 for evaluation, 25 for explanation, and 18 for self-regulation (44% to 62%). These item-level percentages were not accompanied by variability estimates or individual-level item responses. Hence, they reveal a descriptive ordering, not statistically established differences in sub-skill responsiveness. The source additionally reported cross-setting implementation scores of Pr = 3.4–3.6 and prior-achievement-stratified effect-size figures of 1.47 and 0.89. With subgroup definitions, denominators, and uncertainty unavailable, these values cannot substantiate a formal moderation or equity effect.

Patterns Across Development Phases and Learning Artefacts

Six theoretically derived phase-indicator pairs were obtained through matching the ratings of activities on the phase level with the gains of the selected indicators. Investigating was found to have an activity rating of 3.8 with the inference indicator gain of 38 percentage points, while evaluating had an activity rating of 2.9 with the self-regulation indicator gain of 18 percentage points. According to the original study, Spearman's rho for these six points was equal to .91 (p = .012). The ecological correlation between the points does not allow it to be used as an estimate of the predictive relationship between the variables, and the p-value should not be used in order to support phase-related causal claims because the six points reflect different instructional phases and different assessment items.

Five theoretically selected indicator changes were matched with the corresponding changes in artefacts: model book interpretation, teacher guide evaluation/explanation, student workbook inference/analysis, AR application inference, assessment instrument self-regulation. As two artefacts have been mapped to multiple indicators, the technique of aggregating the changes in indicators has not been clearly described. Pearson's r = .88 (p = .048) for five aggregated points according to the original study. This exploratory value is highly mapping-dependent, cannot be treated as evidence that increasing an artefact's rating causes learning gains, and should not be used as a generalizable predictive model. This manuscript retains the numerical result only to document the original exploratory analysis and explicitly limits its interpretation.

Discussion

MathVerse M6 must be evaluated in terms of three types of evidence: usable instructional product evidence from educational design research, mathematics learning evidence from the field of mathematics education on the growth of reasoning, and immersive learning evidence from AR research. While the test results were positive regarding the evaluation of the five artefacts by experts (mean ratings 3.48–3.82 on a scale of four), the high level of implementation in the large-scale test (Pr = 3.5), and the significant difference between pre-test and post-test mathematical critical thinking scores (48 to 82), it becomes a more challenging question of how a structured six-phase pedagogy can convert visual engagement with geometric solids into mathematical reasoning. The following discussion presents a comparison of the observed MathVerse M6 patterns with existing experimental studies, systematic reviews, and metacognitive research.

Product Quality and Learning Gains: The Difference Between Feasibility and Effectiveness (RQ1)

Evidence of coherence and acceptability in MathVerse M6 is shown by the alignment of expert ratings, classroom implementation, and positive student reactions to the learning experience. It also demonstrates one of the key strengths of the design-research approach because the intervention includes more than just the application. The model book, teacher's guide, student workbook, immersive environment, and the assessment tool were developed as a learning system where technological elements were aligned with teacher and learner actions. However, MathVerse M6 differs from research studies that consider the availability of AR devices as an intervention. Nonetheless, while expert ratings represent the quality of designed artifacts and practicality refers to teacher enactment, learning scores represent student achievements. These three are aligned positively but cannot prove that the AR-metaverse element created the effect of a 34-point change on its own.

The overall direction of change is compatible with the wider evidence that well-designed AR can support academic learning. Chang et al. (2022), in a meta-analysis of (quasi-)experimental educational studies, reported positive average effects of AR but also identified treatment duration and the design of three-dimensional visualisation as important sources of variation. More recently, Hidayat et al. (2026) synthesised 22 STEM studies involving 2,215 participants and reported favourable pooled outcomes for academic achievement and thinking skills. These syntheses make the MathVerse M6 improvement plausible as an educational phenomenon; they do not establish that the improvement was caused by MathVerse M6. Unlike effect estimates derived from comparisons between intervention and comparison conditions, the present 48-to-82 change comes from the same pupils tested twice, with no parallel control condition. Directly ranking this change or the disputed standardised effect-size figure against a meta-analytic treatment effect would therefore conflate different estimands.

An even more illuminating contrast comes from studies that examine when immersive technology does not confer an advantage. In their review of 36 studies, Poupard et al. (2025) found that virtual reality could introduce extraneous cognitive load, whereas AR more often supported novice learners by externalising complex content; gains in motivation were less conclusive. Similarly, Coban et al. (2022) reported a modest overall learning effect for immersive virtual reality rather than uniformly large gains. These findings challenge the technological assumption that greater immersion necessarily yields better learning. For MathVerse M6, an alternative explanation for the positive trial profile is the fit between the prism-geometry topic, the structured questioning

in M6, and the possibility of manipulating representations. The relative contribution of each component remains unknown because teacher guidance, task sequencing, AR features, and testing occurred as a single package. Future comparison conditions should therefore distinguish the added value of AR from that of the established M6 pedagogy.

Why Inference and Analysis Appeared to Improve More Than Self-Regulation (RQ2)

The item-linked pattern is particularly worthy of discussion: inference increased by 38 percentage points and analysis by 35, whereas self-regulation increased by 18. The result is descriptively consistent with the intellectual demands of Investigating and Presenting. During investigation, learners can manipulate a geometric solid, observe changes, formulate a conjecture, and compare it with measurements or calculated quantities. During presentation, learners must translate these observations into an explanation that other pupils can question. In this sense, the immersive representation is not the thinking process itself; it becomes educationally useful when it supplies evidence for inference or material for argument. The finding broadly complements Koparan et al. (2023), who studied AR-supported mathematics instruction, and Walkington et al. (2024), who documented how AR dynamic geometry environments provide distinctive opportunities to change perspective, scale, and orientation while reasoning about shapes.

However, the literature also reveals why visible engagement with three-dimensional objects should not be mistaken for deeper mathematical justification. In a preregistered study involving 120 high-school students, Walkington et al. (2025) found that head-mounted AR was especially useful for gaining an initial sense of three-dimensional forms, but it did not consistently improve the justification or generalisation of geometric conjectures relative to tablet-based work. That contrast is consequential for MathVerse M6: a student may rotate a triangular prism, recognise its surfaces, and correctly infer a solution without necessarily being able to articulate why a general relationship holds. Thus, the larger inference and analysis item changes may partly reflect task–representation alignment rather than an established phase-specific cognitive mechanism. A compelling follow-up would separate questions requiring visual identification, inference from virtual experiments, evaluation of an incorrect proof, and general justification. Such differentiation would show whether the intervention supports mathematical argument beyond recognition and manipulation.

Wu et al. (2024) offer another comparison of a different methodology but with a complementary finding. The results of their quasi-experiment on AR math picture books indicate better performance in terms of geometric thinking and flow, as well as lower cognitive load, for fourth graders who used AR compared to the group of students who used conventional materials. This study confirms the viability of the use of visualizations along with mathematics exercises but provides more reliable comparisons between instructional approaches due to its comparison groups. Additionally, geometric thinking, critical-thinking sub-skills, and student flow are different constructs that may overlap but are not equivalent. The contribution of MathVerse M6 is thus the six-indicator diagnostic framework, and not a proof of the specific development of each Facione sub-skill at a particular M6 stage. The sequential development pattern remains, therefore, a reasonable hypothesis because the order of gains cannot be validated due to a single item per indicator and the same experience of all six stages by all students.

The Reflection Bottleneck: From Manipulating Objects to Monitoring One's Reasoning

The relatively small self-regulation item change is not merely a weaker result; it raises a pedagogically productive question about what immersive mathematics environments leave implicit. Students can be highly occupied with exploring a virtual object while engaging only briefly in checking assumptions, identifying errors, or revising an earlier claim. In the present summary, Evaluating also had the lowest recorded phase-activity rating (Ar = 2.9), whereas Investigating had the highest (Ar = 3.8). These two observations cannot prove that low activity caused the smaller self-regulation improvement, since activity and gains were aggregated in different ways. Nevertheless, their convergence offers a practical target for iterative design: Phase 5 may need more explicit metacognitive tasks rather than merely additional digital interaction.

The distinction is reinforced by Mannan et al. (2026), whose systematic review of 18 studies of GeoGebra-supported mathematical problem-solving identifies awareness, regulation, evaluation,

and the gradual appropriation of digital tools as distinct aspects of metacognitive activity. Their synthesis indicates that dynamic visualisation can make a task accessible and facilitate monitoring, but evaluation frequently remains at the level of visual or cross-representational checking. That insight helps explain why visually successful reasoning does not automatically yield substantial self-regulation gains. Pereles et al. (2024) likewise foreground the role of explicit metacognitive strategies in technology-supported critical thinking, while Hwang et al. (2024) examine digital scaffolding designed to support students' self-regulated learning. The common design implication is not to increase immersion indiscriminately, but to make reflection and revision visible as required mathematical actions.

For instance, during the Evaluating stage, learners could be required to save their first solution, examine an intentionally deceptive prism net or erroneous volume calculation, determine the assumption that caused the error, and provide a new rationale. This comparison can reflect on monitoring and evaluating better than a satisfaction rate would. Subsequently, an activity assessing delayed transfer can test whether learners apply the same checking approach without the assistance of AR. This would be a suggestion for further research and not a task grounded in current evidence. It will also aid in distinguishing metacognitive self-regulation from post-task verification and deal with the issue arising from the use of one item for assessing a broad construct.

Prior Achievement, Cognitive Load, and the Question of Educational Equity (RQ3)

As compared to the original subgroup summary, larger standardised changes were observed for lower prior achieving students (1.47) than for higher prior achieving students (0.89). It is likely that a possible explanation is the fact that manipulation can offer an external scaffold to those learners who lack a well-established internal representation of 3D shapes. This idea is aligned with the findings of Poupard et al. (2025), in which it was observed that AR was more helpful to beginners, and research showing that technology-enabled scaffolds can vary based on the different requirements of learners (Kim et al., 2024). In prisms geometry, a novice learner who has difficulties with transforming a net into a solid will find it useful to observe such a transformation process directly. However, the current manuscript does not provide subgroup thresholds, subgroup sizes, confidence intervals, or interaction between time and achievement. Therefore, the reported values suggest only exploratory comparison but not compensatory or equity effects.

Further research explains why the direction of such a contrast would be contingent upon the nature of the task performed. For instance, Walkington et al. (2026), in their preregistered experiment with 70 secondary school students, showed that there were differences between collaboration forms in AR geometry, in which case parallel interaction proved to be superior overall and joint interaction potentially suitable for those having more previous knowledge. Similarly, Walkington et al. (2025) concluded that the positive aspects of alternative visual interfaces were task-specific, depending on whether the shape was two- or three-dimensional and what kind of reasoning was involved. Such conclusions suggest that AR cannot be viewed as a universal equalizer since while one group may benefit from an approach reducing the initial spatial load, another group may face additional challenges with coordinating actions or navigation. In MathVerse M6, the supposed higher performance of the low-achievement group may have been determined by the nature of the prism problems or potential room for score improvement.

Implementation consistency within the reported settings (Pr = 3.4–3.6) is also less evidence than equitable access. Practical lessons may well require students to have differential access to the right equipment, consistent connectivity, technical support, or chances to interact with the technology instead of observing their fellow students. Thus, TPACK becomes theoretically relevant not only in terms of the fit between subject content and the features of the technology but also the school context and available resources of teachers (Brianza et al., 2024; Schmid et al., 2024). A better-designed equity-oriented analysis would need to document the availability of the equipment, previous experience with AR, required support, interaction opportunities, and check for persistence of the differences after accounting for initial achievement and school/class clustering. Without this information, it is legitimate to conclude that implementation seemed possible in the studied schools with an interesting descriptive difference by achievement.

What Phase Activity Can and Cannot Explain About Critical Thinking (RQ4)

On first glance, the Spearman correlation that was found between six phase activities and the six sub-skill changes associated with them (rho = .91) might seem to be evidence for an engagement mechanism. However, the correlation is between six different instructional phases that are also linked to six different measurement variables. It is not a correlation between individual engagement and individual learning gains, and it varies depending on how the mapping was done. It is better to consider this a design heuristic – namely that the highest activity rating corresponded to Investigating with a large change in the number of inference items, whereas the lowest activity rating corresponded to Evaluating with a small change in the number of self-regulation items.

The difference between the two types of participation is important since studies show that all participation is not necessarily as beneficial. Specifically, in the preregistered AR collaboration experiment conducted by Walkington et al. (2026), various gestures and actions occurred depending on the format of interaction, but they did not prove to be positively correlated with test results. Thus, a simplistic more-action–more-learning correlation does not seem to hold. Additionally, Won et al. (2023) point out various ways in which learners interact with immersive environments, while Wang et al. (2024) stress the significance of an educational environment that supports autonomy, competence, and relatedness. In the case of MathVerse M6, the crucial element should be whether students notice something mathematically relevant, make justified conjectures, argue against previous claims, correct mistakes, and explain their reasoning rather than how interesting a particular virtual phase seems. Future LPAS data can include tool navigation, excitement unrelated to learning, collaborative communication, efforts at justifying conclusions, and metacognition checks.

Artefact Validity as a Design Diagnostic Rather Than a Learning Predictor (RQ5)

The correlation between five artefact rating averages and their gains on the theoretically mapped indicators (r = .88) provides yet another potentially valid but scientifically weak explanation for observed results. The expert ratings of a model book, a teacher guide, a workbook, an application, and an instrument are assessments of five different artefacts; test item gains reflect the performance of students. Since some of the artefacts were attributed to several skills, there exists no specific, validated correlation between the quality of the artefact and its gain. Given the small number of points – five – the correlation value can hardly be interpreted as proof that better validating an artefact will enhance critical thinking skills.

But there is an actionable lesson from the validation analysis. The application had an average score of 3.48 across all three aspects while scoring 3.25 on the visualisation aspect, which is less than its scores on the access and content aspects. It must be noted that these ratings do not establish the connection between the visualisation and students' learning problem, but they could help pose formative questions. Is the geometrical accuracy of the virtual prism nets correct? Can the learners observe the hidden faces and their measurements without excessive navigation through the interface? Does the visualisation emphasise the connections that need to be explained? From reviews on multimedia design and immersive learning, we know that there is a danger of distraction by excessive visual details and the necessity to align representational features with pedagogical goals (Çeken & Taşkın, 2025; Poupard et al., 2025).

Theoretical Contribution: Bridging M6, TPACK, and Critical-Thinking Processes

The strongest contribution of MathVerse M6 is not the invention of a wholly new critical-thinking theory or evidence that the metaverse outperforms ordinary classroom teaching. Rather, it is an explicit design argument connecting three levels that are often discussed separately: what mathematics students must reason about (spatial properties of prisms), how teachers organise that reasoning (the six M6 phases), and which technological affordances can make particular representations available at useful moments. The original M6 model already emphasised structured critical thinking (Susandi et al., 2022, 2026); the current contribution extends its enactment through AR-supported activities and proposes an indicator-oriented account of how different classroom episodes might create opportunities for interpretation, inference, evaluation, explanation, and self-regulation.

The position also makes a more accurate application of TPACK possible. As pointed out by Schmid et al. (2024), TPACK can be regarded as an advanced, situated, and dynamic framework, although there are issues with its operationalisation and measurement. In this research, TPACK serves as a rationale for selecting and ordering certain technologies in a meaningful way, rather than a measured variable. Hence, an assertion about the effect of TPACK alignment on the score difference would go beyond what was actually measured. This research brings into focus a tangible instance of an alignment between technology, pedagogy, and content: virtual manipulation occurs in the context of teacher-led questioning, while investigation, presentation, and evaluation have specific cognitive roles. Whether such alignment has any learning benefits can be studied through the comparison of MathVerse M6 and M6 not accompanied by AR technology.

One theoretically intriguing possibility is an asymmetry in the ease with which geometric relationships can be made perceptible vs. evaluative reasoning made explicit. This is supported by the difference in changes in the inference/analysis item vs. the self-regulation item in the trial, and there are independent grounds for considering it plausible due to the controlled AR experiments of Walkington et al. (2025) and the metacognitive synthesis of Mannan et al. (2026). However, the present results do not show any general ranking of all six sub-skills in terms of difficulty. Instead, they inspire the falsifiable hypothesis that representation-focused scaffolding may be particularly effective for spatial inference without explicit reflection and justification. This is a more concrete research path than the general one concerning critical thinking and immersive technology.

Practical Implications, Methodological Boundaries, and Future Research

The implications of the results from this study for teachers would be that they implement a balanced approach to MathVerse. Focusing and Justifying would enable students to provide prior knowledge and explain how a potential formula works; Investigating would enable the use of manipulatives to explore conjectures; Presenting would convert observations into mathematical proofs; Evaluating would involve having students analyse their mistakes in written form; and Concluding would allow the proofs to become linked to a generalised formula. The order of tasks is more important than any display that may look impressive. It may even be as important as designing better three-dimensional images to develop simple prompts to predict, justify, and correct errors. For software developers, the focus must be on accurate geometry, easy-to-use controls, clear feedback, and the ability to export learner reasoning, rather than features of immersion without educational uses.

A number of limitations determine the scope of the interpretation of the findings. The study included intact classrooms in a purposive sample, employed a one-group pre-test/post-test design, presented only one item per critical thinking construct, and did not provide measures of paired-score precision or reliability adequate for identifying subskills. The original standardized effect size is not consistent with the descriptive statistics presented in the paper. The correlations between phases and artifacts are calculated on the basis of six and five total points, respectively, while students' assessments are clustered within the classroom and school levels. This limits causal attribution to AR, independent estimation of phase effects, subgroup moderations, and individual-level predictions. In addition, the study presents a positive reaction of students but does not measure the cognitive load, motivation, or interaction processes referred to as possible mechanisms.

A rigorous next cycle might entail three interrelated phases. First, iteratively develop Phase 5 and document every change according to formative feedback from student solutions, instructor observations, and accessibility constraints as identified in expert rating. Second, test M6 in comparison to M6 without AR through parallel classes or matched comparison groups, pre-register the primary outcome measures, and use multiple calibrated tasks for each Facione indicator and blinded scoring with agreement between raters reported. Third, gather phase-specific process data and follow students' post-intervention, enabling clustered and multilevel analysis of retention, transfer, and heterogeneous impact of learning as a function of initial achievement and accessibility condition. This research agenda fits the calls in the general literature to go beyond novelty and average gain to an explanation of the conditions under which particular learners benefit through particular mathematical activities (Chang et al., 2022; Poupard et al., 2025; Walkington et al., 2025,

2026). Therefore, MathVerse M6 should be viewed as an innovative, theory-based design, which initial empirical evidence suggests requires systematic refinement and replication.

LIMITATIONS

There are a number of limitations associated with this study. First of all, this research involved purposeful sampling, with 16 teachers and 118 students in Grade 9 from four different schools in Cirebon, Indonesia. Secondly, the large-scale pilot experiment had a one-group pre-test/post-test design, meaning that there was no control group used; hence, the degree to which improvement in student scores is related to MathVerse M6 cannot be assessed fully. In addition, the small-scale trial did not include a corresponding pre-test, so changes in students' performance during this stage could not be assessed. Third, the critical-thinking test comprised six items, with one item representing each of the six indicators. Consequently, the reported indicator-level results should be interpreted in light of the limited number of items used to assess each indicator. Further studies involving more diverse educational settings, comparative research designs, and a broader set of assessment items are needed to examine the applicability of MathVerse M6 and further investigate students' critical-thinking development.

CONCLUSION

MathVerse M6 combined an established six-phase mathematics learning model with an AR-supported virtual environment and associated instructional materials. Across five artefacts, expert mean ratings were 3.48–3.82 on a four-point scale, and large-scale implementation received a practicality rating of 3.5. Among the 90 students assessed before and after implementation, the recorded mean test score increased from 48 to 82; the proportion attaining at least 75 increased from 0/90 to 75/90. Item-level summaries indicated the largest descriptive changes for inference and analysis and the smallest for self-regulation. The reported phase–indicator and artefact–indicator correlations generate hypotheses but are based on six and five aggregate observations, respectively, and cannot identify causal mechanisms or student-level prediction. The one-group design, ambiguous standardized effect size data, cluster groups, and insufficient sub-skill data render the results formative and exploratory in nature. MathVerse M6 merits further evaluation with controlled comparisons, more reliable multi-item assessment, explicit design-iteration documentation, and appropriate clustered-data analysis.

AUTHOR CONTRIBUTIONS

ADS was involved in the concept and design of the study, in formulating the research design, and writing the initial draft of the manuscript. PD and IR were involved in the methodology, in analysing the mathematics education-related aspects and interpretation of results. SF was involved in the formulation of the research idea and in the critical review of the manuscript. TK was involved in the development of the framework for mathematics learning and interpretation of the educational aspects. EP was involved in the pedagogic aspects of the research and in the critical review of the manuscript. HB was involved in the interpretation of the mathematics education-related aspects and in reviewing the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the Universitas Terbuka for funding under contract number B/978/UN31.LPPM/PT.01.03/2025. This work was supported by the Indonesian Collaborative Research Grant (RKI) Scheme A 2025, involving Universitas Terbuka, Universitas Negeri Malang, and Universitas Pendidikan Indonesia.

REFERENCES

Alfaisal, R., Hashim, H., & Azizan, U. H. (2023). Metaverse system adoption in education: A systematic literature review. *Journal of Computers in Education*. https://doi.org/10.1007/s40692-022-00256-6

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*, 462–486. https://doi.org/10.1007/s40751-024-00152-x

Brianza, E., Schmid, M., Tondeur, J., & Petko, D. (2024). Is contextual knowledge a key component of expertise for teaching with technology? A systematic literature review. *Computers and Education Open, 7*, Article 100201. https://doi.org/10.1016/j.caeo.2024.100201

Bulut, M., & Borromeo Ferri, R. (2023). A systematic literature review on augmented reality in mathematics education. *European Journal of Science and Mathematics Education, 11*(3), 556–572. https://doi.org/10.30935/scimath/13124

Çakıroğlu, Ü., Güler, M., Dündar, M., & Coşkun, F. (2023). Virtual reality in realistic mathematics education to develop mathematical literacy skills. *International Journal of Human–Computer Interaction, 40*(17), 4661–4673. https://doi.org/10.1080/10447318.2023.2219960

Camilleri, M. A. (2024). Metaverse applications in education: A systematic review and a cost-benefit analysis. *Interactive Technology and Smart Education, 21*(2), 245–269. https://doi.org/10.1108/ITSE-01-2023-0017

Canbaz, B., & Yalçın, N. (2024). The effect of mathematics teaching with mobile augmented reality technology on secondary school students' attitudes and academic achievements. *Participatory Educational Research, 11*(4), 59–76. https://doi.org/10.17275/per.24.49.11.4

Çeken, B., & Taşkın, N. (2025). Examination of multimedia learning principles in augmented reality and virtual reality learning environments. *Journal of Computer Assisted Learning, 41*(1), Article e13097. https://doi.org/10.1111/jcal.13097

Chamola, V., Peelam, M. S., Mittal, U., Hassija, V., Singh, A., Pareek, R., Mangal, P., Sangwan, D., de Albuquerque, V. H. C., Mahmud, M., & Brown, D. J. (2025). Metaverse for education: Developments, challenges, and future direction. *Computer Applications in Engineering Education, 33*(3), Article e70018. https://doi.org/10.1002/cae.70018

Chang, H.-Y., Binali, T., Liang, J.-C., Chiou, G.-L., Cheng, K.-H., Lee, S. W.-Y., & Tsai, C.-C. (2022). Ten years of augmented reality in education: A meta-analysis of (quasi-) experimental studies to investigate the impact. *Computers & Education, 191*, Article 104641. https://doi.org/10.1016/j.compedu.2022.104641

Coban, M., Bolat, Y. I., & Goksu, I. (2022). The potential of immersive virtual reality to enhance learning: A meta-analysis. *Educational Research Review, 36*, Article 100452. https://doi.org/10.1016/j.edurev.2022.100452

De Felice, F., Petrillo, A., Iovine, G., Salzano, C., & Baffo, I. (2023). How does the metaverse shape education? A systematic literature review. *Applied Sciences, 13*(9), Article 5682. https://doi.org/10.3390/app13095682

Erşen, Z. B., & Alp, Y. (2024). A systematic review of augmented reality in mathematics education in the last decade. *Malaysian Online Journal of Educational Technology, 12*(1), 15–31. https://doi.org/10.52380/mojet.2024.12.1.476

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report).* American Philosophical Association.

Feng, S., Xu, X., Li, S., Li, Z., & Gibson, D. (2024). Is metaverse a buzzword in education? Insights from a systematic review. *Educational Technology Research and Development, 72*, 3349–3390. https://doi.org/10.1007/s11423-024-10398-2

Gómez-Rodríguez, V. (2026). From virtual worlds to real-world equity: A scoping review of the metaverse as computer-assisted learning for STEM competencies. *Computers, 15*(4), Article 229. https://doi.org/10.3390/computers15040229

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), Article e13095. https://doi.org/10.1111/jcal.13095

Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2021). Immersive virtual reality as a pedagogical tool in education: A systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education, 8*, 1–32. https://doi.org/10.1007/s40692-020-00169-2

Hidajat, F. A. (2024). Augmented reality applications for mathematical creativity: A systematic review. *Journal of Computers in Education, 11*, 991–1040. https://doi.org/10.1007/s40692-023-00287-7

Hidayat, R., Samsudin, S., Wang, X., & Abd Rahman, S. N. (2026). Augmented reality in science, technology, engineering and mathematics education: A meta-analysis. *Journal of Computer Assisted Learning, 42*(1), Article e70164. https://doi.org/10.1002/jcal.70164

Hwang, G.-J., Chang, C.-Y., & Chang, S.-C. (2024). Promoting students' self-regulated learning with AI-incorporated metacognitive scaffolding in a flipped learning context. *British Journal of Educational Technology, 55*(1), 210–228. https://doi.org/10.1111/bjet.13384

Hwang, G.-J., Tu, Y.-F., & Chu, H.-C. (2023). Conceptions of the metaverse in higher education: A draw-a-picture analysis and surveys to investigate the perceptions of students with different motivation levels. *Computers & Education, 203*, Article 104868. https://doi.org/10.1016/j.compedu.2023.104868

İslim, Ö. F., Namlı, Ş., Sevim, N., Özçakır, B., & Lavicza, Z. (2024). Augmented reality in mathematics education: A systematic review. *Participatory Educational Research, 11*(4), 115–139. https://doi.org/10.17275/per.24.52.11.4

Istiqlal, M., Retnawati, H., Widihastuti, & Sari, D. K. (2026). Assessing students' computational thinking ability based on mathematical reasoning. *Jurnal Ilmiah Peuradeun, 14*(2), 773–796. https://doi.org/10.26811/peuradeun.v14i2.1602

Kim, N. J., Belland, B. R., & Axelrod, D. (2024). Adaptive scaffolding, technology, and mathematics learning: A systematic review of the equity implications. *Educational Technology Research and Development, 72*(1), 89–118. https://doi.org/10.1007/s11423-023-10285-6

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*, Article 101245. https://doi.org/10.1016/j.tsc.2023.101245

Ku, K. Y. L., Kong, Q., Song, Y., Hou, L., Deng, J., & Li, X. (2024). What predicts critical thinking performance? Investigating the role of prior knowledge, epistemological beliefs, and metacognitive self-regulation. *Thinking Skills and Creativity, 51*, Article 101481. https://doi.org/10.1016/j.tsc.2024.101481

Leppink, J. (2024). Cognitive load theory: Advances and contemporary perspectives. *Educational Psychology Review, 36*(1), Article 17. https://doi.org/10.1007/s10648-023-09817-6

Li, M., & Yu, Z. (2023). A systematic review on metaverse-based blended English learning. *Frontiers in Psychology, 13*, Article 1087508. https://doi.org/10.3389/fpsyg.2022.1087508

Maghaydah, S., Al-Emran, M., Maheshwari, P., & Al-Sharafi, M. A. (2024). Factors affecting metaverse adoption in education: A systematic review, adoption framework, and future research agenda. *Heliyon, 10*(7), Article e28602. https://doi.org/10.1016/j.heliyon 2024.e28602

Mannan, A. M. K., Bajuri, M. R., Abdul Rahim, S. S., & Abdullah, N. A. (2026). How do digital tools support metacognition in mathematics learning? A systematic review of research on GeoGebra and problem-solving. *Educational Research Review, 52*, Article 100827. https://doi.org/10.1016/j.edurev.2026.100827

Maulida, A. S., Wahyudin, Turmudi, & Nurlaelah, E. (2024). The effect of experiential learning and directed instruction assisted by augmented reality on students' self-regulated learning. *Infinity Journal, 13*(2), 553–568. https://doi.org/10.22460/infinity.v13i2.p553-568

Miller, M. D., Kong, A., & Oh, J. (2024). Edu-metaverse design: Perspectives of undergraduate learners. *Computers & Education: X Reality, 5*, Article 100079. https://doi.org/10.1016/j.cexr.2024.100079

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Mishra, P., Warr, M., & Islam, R. (2023). TPACK in the age of ChatGPT and generative AI. *Journal of Digital Learning in Teacher Education, 39*(4), 235–251. https://doi.org/10.1080/21532974.2023.2247480

Nieveen, N., & Folmer, E. (2013). Formative evaluation in educational design research. In T. Plomp & N. Nieveen (Eds.), *Educational design research* (pp. 152–169). SLO Netherlands Institute for Curriculum Development.

Onu, P., Pradhan, A., & Mbohwa, C. (2024). Potential to use metaverse for future teaching and learning. *Education and Information Technologies, 29*, 8893–8924. https://doi.org/10.1007/s10639-023-12167-9

Pereles, A., Ortega-Ruipérez, B., & Lázaro, M. (2024). The power of metacognitive strategies to enhance critical thinking in online learning. *Journal of Technology and Science Education, 14*(3), 831–843. https://doi.org/10.3926/jotse.2721

Plomp, T. (2013). Educational design research: An introduction. In T. Plomp & N. Nieveen (Eds.), *Educational design research* (pp. 11–50). SLO Netherlands Institute for Curriculum Development.

Poupard, M., Larrue, F., Sauzéon, H., & Tricot, A. (2025). A systematic review of immersive technologies for education: Learning performance, cognitive load and intrinsic motivation. *British Journal of Educational Technology, 56*(1), 5–41. https://doi.org/10.1111/bjet.13503

Pradana, M., & Elisa, H. P. (2023). Metaverse in education: A systematic literature review. *Cogent Social Sciences, 9*(2), Article 2252656. https://doi.org/10.1080/23311886.2023.2252656

Ramadoni, & Adnan, M. (2026). Cross-national effects of a gamified algebra learning intervention: A mixed-design analysis of country and gender differences in Indonesia and Malaysia. *Eurasia Journal of Mathematics, Science and Technology Education, 22*(5), Article em2829. https://doi.org/10.29333/ejmste/16199

Risdianto, E., Wachidi, W., Riyanto, R., Fathurochman, I., Yanto, M., & Asmara, A. (2022). Feasibility test of learning media with blended learning model and augmented reality-assisted MOOCs. *Jurnal Ilmiah Peuradeun, 10*(1), 149–164. https://doi.org/10.26811/peuradeun.v10i1.626

Romero-Saritama, J. M., Espinosa-Tapia, S., & Palacios-Rodríguez, A. (2025). Using WhatsApp in distance education: Assessing the impact on academic interaction and influencing factors. *Social Sciences, 14*(3), Article 183. https://doi.org/10.3390/socsci14030183

Sakr, A., & Abdullah, T. (2024). Virtual, augmented reality and learning analytics impact on learners, and educators: A systematic review. *Education and Information Technologies, 29*, 19913–19962. https://doi.org/10.1007/s10639-024-12602-5

Schmid, M., Brianza, E., Mok, S. Y., & Petko, D. (2024). Running in circles: A systematic review of reviews on technological pedagogical content knowledge (TPACK). *Computers & Education, 214*, Article 105024. https://doi.org/10.1016/j.compedu.2024.105024

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE.

Soebagyo, J., Usman, H., & Tambak, S. (2023). Digital divide and technology integration in Indonesian mathematics classrooms: A nationwide survey. *Journal on Mathematics Education, 14*(4), 611–632. https://doi.org/10.22342/jme.v14i4.pp611-632

Susandi, A. D., Sa'dijah, C., As'ari, A. R., & Susiswo. (2022). Developing the M6 learning model to improve mathematical critical thinking skills. *Pedagogika, 145*(1), 182–204. https://doi.org/10.15823/p.2022.145.11

Susandi, A. D., Sa'dijah, C., & Susiswo, S. (2026). Improving critical thinking skills in mathematics: Implementation of the M6 learning model. *Pedagogical Research, 11*(3), Article em0271. https://doi.org/10.29333/pr/18773

Umam, K., & Susandi, A. D. (2022). Critical thinking skills: Error identifications on students with APOS theory. *International Journal of Evaluation and Research in Education, 11*(1), 182–192. https://doi.org/10.11591/ijere.v11i1.22048

Vegas, S. O., & Djukri. (2021). The effect of problem-based learning on the critical thinking skill of the students. *Jurnal Ilmiah Peuradeun, 9*(1), 221–240. https://doi.org/10.26811/peuradeun.v9i1.411

Walkington, C., Nathan, M. J., Hunnicutt, J., Washington, J., & Zhou, M. (2024). New kinds of embodied interactions that arise in augmented reality dynamic geometry software. *The Journal of Mathematical Behavior, 75*, Article 101175. https://doi.org/10.1016/j.jmathb.2024.101175

Walkington, C., Nathan, M. J., Washington, J., Hunnicutt, J., Darwin, T., Daughrity, L., & Schenck, K. (2025). Comparing learning geometry using a tablet to head-mounted display augmented reality: How and when dimensionality matters. *Education and Information Technologies, 30*, 5397–5426. https://doi.org/10.1007/s10639-024-13008-z

Walkington, C., Washington-Henderson, J., Hunnicutt, J., & Nathan, M. J. (2026). The impact of different collaboration formats on mathematical problem-solving in augmented reality. *Computers & Education, 242*, Article 105491. https://doi.org/10.1016/j.compedu.2025.105491

Wang, Q., & Abdullah, A. H. (2024). Enhancing students' critical thinking through mathematics in higher education: A systematic review. *SAGE Open, 14*(3), 1–15. https://doi.org/10.1177/21582440241274007

Wang, X., Liu, Q., Pang, H., Lv, Z., Sun, D., Chen, W., & Cheng, B. (2023). What matters in the design of game-based learning for mathematics? A meta-analysis on learning achievement and motivation in China. *Computers & Education, 195*, Article 104718. https://doi.org/10.1016/j.compedu.2023.104718

Wang, Y., Wang, H., Wang, S., Wind, S. A., & Gill, C. (2024). A systematic review and meta-analysis of self-determination-theory-based interventions in the education context. *Learning and Motivation, 87*, Article 102015. https://doi.org/10.1016/j.lmot.2024.102015

Won, M., Ungu, D. A. K., Matovu, H., Treagust, D. F., Tsai, C.-C., Park, J., Mocerino, M., & Tasker, R. (2023). Diverse approaches to learning with immersive virtual reality identified from a systematic review. *Computers & Education, 195*, Article 104701. https://doi.org/10.1016/j.compedu.2022.104701

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*, 24627–24652. https://doi.org/10.1007/s10639-024-12768-y

Xie, J., Al-Shaibani, G. K. S., Zhao, J., & Talib, O. (2025). The current status and contributing factors of the metaverse adoption in education: A systematic review. *Contemporary Educational Technology, 17*(4), Article ep593. https://doi.org/10.30935/cedtech/17301

Zapata-Rodríguez, M., Ramos-Galarza, C., Valencia-Aragón, K., & Guachi, L. (2024). Enhancing mathematics learning with 3D augmented reality escape room. *International Journal of Educational Research Open, 7*, Article 100389. https://doi.org/10.1016/j.ijedro.2024.100389