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.

How teachers refine and enact culturally situated STEAM worksheets drafted by generative AI for learning matrices in lesson study

Muhamad Galang Isnawan
Universitas Nahdlatul Wathan Mataram
Naif Mastoor Alsulami
University of Jeddah
Lalu Sirajul Hadi
Universitas Mataram
Sudirman Sudirman
Indonesia Open University
Samsuriadi Samsuriadi
Universitas Nahdlatul Wathan Mataram
R. Didi Kuswara
Universitas Nahdlatul Wathan Mataram
I Ketut Sukarma
ketutsukarma@undikma.ac.id

Universitas Pendidikan Mandalika

Research Article

Keywords: generative artificial intelligence, design-based research, epistemic learning patterns, ethnomathematics, teacher agency, linear algebra

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11355632/v1

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

Additional Declarations: No competing interests reported.

How teachers refine and enact culturally situated STEAM worksheets drafted by generative AI for learning matrices in lesson study

Muhamad Galang Isnawan[1], Naif Mastoor Alsulami[2], Lalu Sirajul Hadi[3], Sudirman[4], Samsuriadi[1], R. Didi Kuswara[1], I Ketut Sukarma[5*]

[1]Universitas Nahdlatul Wathan Mataram, Mataram, Indonesia

[2]University of Jeddah, Jeddah, Saudi Arabia

[3]Universitas Mataram, Mataram, Indonesia

[4]Universitas Terbuka, Tangerang Selatan, Indonesia

[5]Universitas Pendidikan Mandalika, Mataram, Indonesia

***Corresponding author:** I Ketut Sukarma, ketutsukarma@undikma.ac.id, +62 818 0522 9595

Email and ORCID

Muhamad Galang Isnawan: galangisna19@gmail.com; https://orcid.org/0000-0002-3799-3435

Naif Mastoor Alsulami: nsulami@uj.edu.sa; https://orcid.org/0000-0002-7404-8141

Lalu Sirajul Hadi: mamiqdamai@staff.unram.ac.id; https://orcid.org/0009-0006-1782-8666

Sudirman: sudirman.official@ecampus.ut.ac.id; https://orcid.org/0000-0002-1696-5160

Samsuriadi: samsuriadimatematika@gmail.com; https://orcid.org/0009-0009-3693-1935

R. Didi Kuswara: r.didi@unwmataram.ac.id; https://orcid.org/0000-0003-0062-4668

I Ketut Sukarma: ketutsukarma@undikma.ac.id; https://orcid.org/0009-0008-0240-702X

Abstract

Generative artificial intelligence (AI) can now draft complete classroom materials, yet little is known about how such drafts fare once teachers refine and use them. This design-based study traced worksheets on matrices from an AI design prompt to classroom enactment in an Indonesian public upper secondary school. The worksheets combined an epistemic learning sequence with science, technology, engineering, arts, and mathematics (STEAM) contexts drawn from local culture. A model teacher taught them to 36 grade 11 students in four lesson study cycles observed by a second teacher and two researchers, and a conventionally taught class (n = 35) took the post-test for comparison. The confirmed worksheets conformed fully to only 6 of 48 coded prompt specifications. The lessons nonetheless sustained active participation, although written justification was often brief and largely missing in the final lesson, misconceptions appeared in two lessons, and unrendered notation had to be repaired by the teacher and by students. All students improved from pre-test to post-test (normalised gain = 0.55; d = 1.93), and the class outscored the comparison class (r = .49). Because the comparison class had no pre-test and some test content overlapped with the worksheets, these results indicate compatibility with learning rather than an effect. The value of AI-drafted materials depended on teachers' decisions during confirmation and enactment.

Keywords generative artificial intelligence, design-based research, epistemic learning patterns, ethnomathematics, teacher agency, linear algebra

1 Introduction

Generative artificial intelligence (AI) has entered classrooms faster than research on its pedagogical consequences. Since the public release of ChatGPT in late 2022, large language models have been promoted as tools that can draft explanations, tasks, and complete lesson materials in seconds, promising relief from teachers' workload and provoking worldwide debate among educators [1, 2]. The same literature warns that

these systems produce fluent but sometimes inaccurate output, and that their educational value depends on the competence of the people who use them [2–4]. Reviews of AI in education have called for greater attention to ethics, collaboration, and methodological rigour [5]. Earlier work on AI for teachers identified lesson planning as a promising use and classroom implementation as a persistent challenge [6]. Generative AI has made this gap more pressing, because teachers can now use these tools directly to produce the materials that students work with.

Reviews report that ChatGPT can raise student engagement when students use it directly [7, 8], but in many classrooms generative AI reaches students indirectly, through materials that teachers prepare with it. In mathematics education, research on AI-generated teaching materials has grown rapidly but remains concentrated on the quality of the output rather than on its use. A scoping survey of ChatGPT in mathematics education identified lesson planning and task design as prominent uses while stressing the need for human oversight [4]. An early rapid review found ChatGPT's performance in mathematics weaker than in several other subject domains [9], although mathematics educators see potential in it for explanation and practice when its limitations are managed [10]. Systematic reviews in mathematics education confirm that ChatGPT can support lesson planning yet struggles with more demanding mathematics [11], and that research on AI in mathematics education is still small and largely qualitative [12]. Preservice secondary teachers who used ChatGPT as a planning assistant found its suggestions often teacher-centred, repetitive, and occasionally mathematically incorrect, although most of them still revised their plans in response to it [13], and analyses of ChatGPT-generated lesson plans stress that teachers must evaluate them critically rather than adopt them as given [14]. Even AI-generated support that produces learning gains can contain frequent errors: in one controlled evaluation, 32% of ChatGPT-generated hints for mathematics problems failed quality checks [15]. Experimental studies of ChatGPT in education have multiplied, and a meta-analysis reports positive effects on academic performance while noting methodological limitations in the primary studies [16]; such syntheses, however, cannot show what teachers changed in AI-generated materials before or during enactment. Such changes are to be expected: teachers interpret and tailor the materials they receive, acting as lesson designers rather than implementers [17], and doing so with generative AI draws on their technological pedagogical content

knowledge [18] and raises new questions about how teachers retain agency over what the tools produce [19]. Consequently, little is known about what happens between an AI draft and the lesson that students actually experience.

Matrices offer a demanding test case. In Indonesian upper secondary schools, matrices are taught in grade 11 as a foundation for later work in linear algebra. Research on algebra learning shows that students' difficulties often centre on symbols, equivalence, and systems of equations, and that teaching practices supporting shared reasoning help students make sense of them [20]. Elementary notions such as element, order, equality, and the properties of matrix operations are therefore not trivial. They require students to interpret notation, justify procedures, and explain why a property holds, not only to compute.

STEAM approaches and culturally grounded mathematics have been proposed to make such abstract content meaningful. STEAM integrates science, technology, engineering, arts, and mathematics, and research on it has grown rapidly while remaining conceptually diverse [21]. Later reviews set the prospects of STEAM alongside persistent problems in how it is defined and implemented [22], and a review of secondary mathematics found only 19 STEAM studies in high-impact journals [23]. Analyses of integrated STEM curricula show that disciplines are often only loosely connected, and distinguish conceptual integration from merely contextual integration [24]. Ethnomathematics research in Indonesia has shown that local cultural practices, such as shadow puppetry and traditional calculations for house building, embed mathematical ideas and values that can anchor school learning and inform the design of didactical situations [25, 26]. In Lombok, the setting of this study, research on the Sasak Wariga calendar has documented modular arithmetic and nested cycles that could support contextual mathematics learning [27]. STEAM+X has been proposed as a way of extending STEAM to further transdisciplinary dimensions [28]. In this study, the "+X" in STEAM+X denotes such local elements, namely culture, vernacular architecture, and local history, added to the STEAM disciplines. Whether AI tools can generate culturally situated tasks without reducing culture to decoration remains an open question.

A further requirement is epistemic. Learning mathematics involves not only producing correct answers but also building, justifying, and communicating knowledge. Research on epistemic cognition describes these activities in terms of the aims, ideals, and reliable processes through which people come to know [29]. Barzilai and Chinn [29] describe epistemic education as promoting apt epistemic performance: reliable reasoning toward valuable epistemic aims, metacognitive regulation of that reasoning, and participation in knowledge construction with others. These aspects point to classroom activities in which students conjecture before being told, test their ideas on problems, explain their solutions to peers, and reflect on what they have come to know.

The Epistemic Learning Patterns (ELP) framework originates in didactical design research conducted through lesson study, where it was introduced as an epistemic lecture pattern for STEM-based courses with prospective mathematics teachers [30]. Its core activities follow the theory of didactical situations, moving learners from action through formulation and validation to institutionalisation, and it closes with self-assessment of both learning and feelings [30]. The present study adapts ELP to secondary mathematics and operationalises the aspects above in three phases. In the epistemic prime phase, students state their readiness, read a contextual story, and make initial conjectures. In the epistemic action phase, they solve a contextual problem in groups and present their solutions for peer response. In the epistemic reflection phase, they formulate conclusions, practise, and reflect on their learning and its personal value. In the worksheets examined here, the phases were realised through a fixed sequence of activities (Let’s Read, Let’s Guess, Let’s Solve the Problem, Let’s Tell, Let’s Conclude, Let’s Practise, and My Reflection) combined with STEAM+X contexts. We refer to the resulting design as ELP-STEAM+X. Our use of the term epistemic draws on, but does not claim to measure, the constructs described by Barzilai and Chinn [29].

Turning such a design into classroom practice requires teacher collaboration, and lesson study provides an established mechanism for it. In lesson study, teachers jointly plan a research lesson, teach and observe it, and reflect on student learning. A review of 75 recent lesson study studies in mathematics education examined how lesson study was implemented and what impact it had on teachers and students [31], and lesson study can help teachers orchestrate whole-class discussions of students’

solutions [32]. Design-based research offers a complementary frame. It develops practical solutions and theoretical insight together, in real settings and in collaboration with practitioners, through iterative cycles of analysis, design, and evaluation [33, 34].

The studies closest to ours combine generative AI with collaborative planning. Cheng [35] followed four science teachers in Hong Kong who used ChatGPT within lesson study to teach density to Grade 7 students, and reported more efficient collaborative planning and improved concept test results. Şimşek [36] analysed how teachers and preservice teachers interacted with ChatGPT when building story-based 5E mathematics lesson plans. Isnawan et al. [30] used lesson study within didactical design research to develop a STEM-based lecture design built on an epistemic learning sequence, and reported gains in prospective teachers' representational abilities alongside problems with time allocation. The present study differs from these in three respects. First, the AI produced complete student worksheets from a structured design prompt rather than lesson plans. Second, the worksheets embedded an explicit epistemic sequence and local cultural contexts. Third, we traced the materials from prompt, through teacher refinement, to classroom enactment and student achievement in mathematics.

Taken together, these strands reveal a gap. Research on generative AI in mathematics education has examined the quality of AI outputs and teachers' planning interactions, and research on lesson study has examined how teachers refine lessons together. How AI-drafted, culturally situated worksheets function once teachers refine and enact them in secondary mathematics classrooms remains scarce, particularly in non-Western settings such as Indonesia. This matters for teachers who increasingly start from AI drafts, and for designers who need principles for when and how teacher judgement must intervene.

This study therefore used design-based research, with lesson study as its participatory mechanism, to design and enact ELP-STEAM+X worksheets on matrices in an Indonesian upper secondary school. It addressed three research questions (RQ1–RQ3):

RQ1: How did teachers and researchers use generative AI to design ELP-STEAM+X worksheets on matrices, and to what extent did the AI outputs conform to the design prompt?

RQ2: How did the worksheets function during classroom enactment across four lesson study cycles?

RQ3: How did students' achievement in matrices change after learning with the worksheets, and how did their post-test scores compare with those of a class taught conventionally?

2 Methods

2.1 Research design

This study followed educational design research [34] and comprised one macro-cycle with three phases: analysis and exploration, design and construction, and evaluation and reflection. Lesson study served as the participatory mechanism within the evaluation phase. Each of the four lessons formed a micro-cycle of plan, do, and see [31]. A quasi-experimental component was embedded to provide supporting evidence on achievement. It combined a one-group pretest–posttest design in the experimental class with a posttest-only comparison against a non-equivalent control class. The design-based framing was chosen because the study aimed to develop and understand a classroom design in situ, not to establish causal effects.

The actual sequence departed from an ideal design-research cycle in one respect: the AI-generated drafts were produced and confirmed during a teacher focus group discussion (FGD), and the raw AI outputs were not retained. The worksheets analysed here are therefore the versions confirmed by teachers and researchers, not the unedited AI output. Table 1 maps each research question to its data and analysis.

Table 1 Research questions, data sources, and analyses

Research Question	Data Sources	Analysis
RQ1: AI-assisted design and conformity to the prompt	Design prompt; four final worksheets; FGD notes	Component-by-component conformity coding
RQ2: Functioning during enactment	Post-lesson reflection notes; photographs of classroom activity; student work written on the interactive board; four completed group worksheets	Qualitative descriptive analysis; coding of worksheet answers
RQ3: Achievement change and comparison	Pre-test and post-test (experimental class); post-test (control class)	Descriptive statistics, paired and independent comparisons, normalised gain, effect sizes

2.2 Setting and participants

The study took place in a public upper secondary school in Mataram, West Nusa Tenggara, Indonesia, between August and September 2026. Two intact grade 11 classes taught by the same teacher were selected purposively. One class was assigned to the ELP-STEAM+X worksheets (experimental class, n = 36; 15 male, 21 female). The other received conventional instruction (control class, n = 35; 15 male, 20 female). Students were aged 17 to 19 years. Random assignment was not possible, and no pre-test was administered to the control class. According to the teacher, the two classes had comparable mathematical ability; this report was not verified by test data.

The lesson study team comprised a model teacher, one observer teacher, and two researchers. Both teachers were male, held a bachelor's degree in mathematics education, and had 15 years of teaching experience. The model teacher taught both classes. In the earlier FGD, 25 mathematics teachers took part in a regional workshop on digital mathematics learning platforms, during which the worksheets were drafted.

2.3 Design prompt and worksheet development

The researchers wrote a structured design prompt that specified every worksheet component. The prompt covered group identity and date; a learning objective stated in one sentence at the analysing, evaluating, or creating level of the revised Bloom taxonomy; an emoticon-based check of students' feelings; an illustrated three-paragraph STEAM+X story ending with a guiding question (Let's Read); prerequisite questions (Let's Guess); an open-ended STEAM+X problem with two solution spaces, the second containing hints (Let's Solve the Problem); a presentation and peer-response task (Let's Tell); a correct solution, concept summary, infographic, and formulas with symbol explanations (Let's Conclude); two open-ended STEAM+X problems, the second at a higher-order level (Let's Practise); and five closing items (My Reflection). The prompt defined X as local culture, vernacular architecture, and local history. It allotted 15, 60, and 15 minutes to the opening, core, and closing activities.

During the FGD, ChatGPT (GPT-5) generated the worksheet content from this prompt, and Gamma.app was used to lay out the slides. Teachers and researchers confirmed the mathematical content. FGD notes recommended further teacher–researcher discussion to validate the worksheets and the tests, and this discussion took place before enactment. The final worksheets were displayed and completed on an interactive flat-panel board during lessons.

2.4 Procedure

The experimental class completed the pre-test before the first lesson. It then took part in four lesson study cycles, held on 15 August, 29 August, 1 September, and 3 September 2026. Each lesson lasted 90 minutes (2 × 45 minutes), and the four lessons covered, in order, the concept, order, and elements of a matrix; types of matrices and the transpose; equality of two matrices; and algebraic operations on matrices. The observer teacher and the researchers observed every lesson, and each lesson was followed by a 15- to 30-minute reflection at the school, whose notes were recorded.

Over the same period, the same teacher taught the control class the same content, in the same number of lessons and with the same time allocation, by explaining the material, working through examples, and assigning textbook exercises. Both classes completed the post-test after the fourth lesson. Figure 1 shows the full procedure. The four lesson study cycles took place only in the experimental class, and the analyses for the three research questions followed data collection.

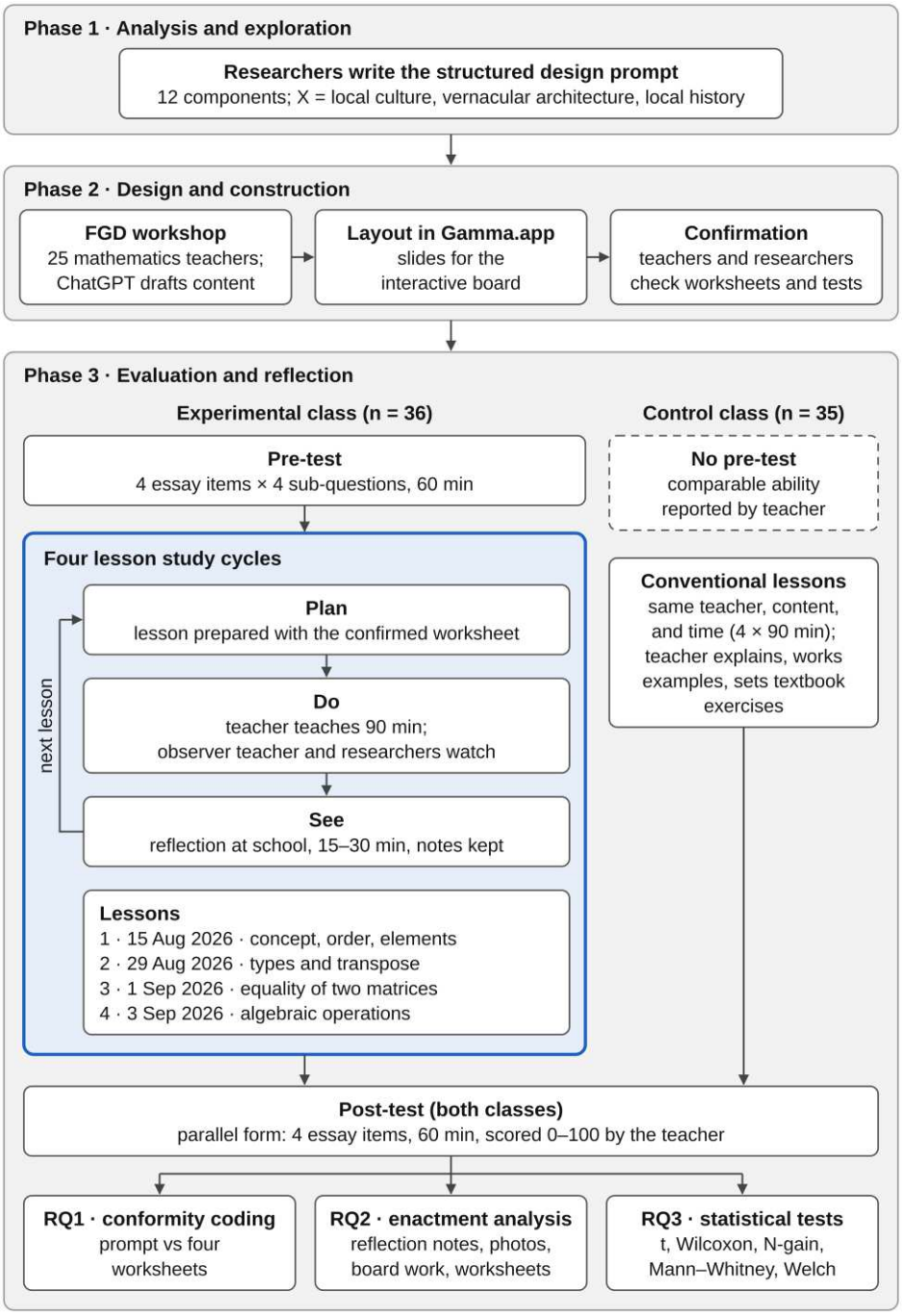

Fig. 1 Research procedure

2.5 Instruments

The pre-test and post-test were parallel essay tests developed by the researchers and teachers. Each contained four items aligned with the four lessons, and each item had four sub-questions (Table 2). Sub-questions a to c required representation or calculation, and sub-question d required an explanation or justification. The two forms used different contexts and numbers. Testing time was 60 minutes.

Table 2 Test blueprint

Item	Content	Pre-test Context	Post-test Context
1	Concept, order, and elements	Extracurricular participation	Library borrowing
2	Types of matrices and transpose	Abstract matrices	Abstract matrices
3	Equality of two matrices	Solving for x and y	Solving for a and b
4	Algebraic operations and their properties	School cooperative sales	School canteen sales

Two features of the instruments affect interpretation. Sub-question 2d differed in focus between forms: the pre-test asked about the role of the identity matrix, whereas the post-test asked about uses of the transpose. In addition, some test content overlapped with worksheet activities. Pre-test item 4 used the same matrices and sales context as the lesson 4 worksheet, and post-test item 1d closely resembled a question in the lesson 1 worksheet.

Each item was worth 25 points, for a maximum of 100. The teacher scored each item with a four-category rubric that credited both the solution procedure and the result (Table 3). Content validity was established through the teacher–researcher discussion recommended in the FGD. No item-level reliability analysis was conducted.

Table 3 Scoring rubric for each test item

Category	Description	Score
1	Procedure and result both correct	25
2	Procedure correct, result incorrect	15
3	Procedure incorrect or incomplete, result correct	10
4	Procedure and result both incorrect, or no answer	0

2.6 Data analysis

RQ1. Each of the four worksheets was compared with the design prompt component by component. Each component was coded as conforming, partially conforming, or deviating. Two researchers coded the worksheets independently and resolved disagreements through discussion until they reached consensus.

RQ2. Reflection notes, photographs, student work visible on the interactive board, and completed group worksheets were analysed through data reduction, display, and conclusion drawing. The analysis focused on student participation, the correctness of written answers, misconceptions, and implementation problems raised in reflection. Completed worksheets were available for one group per lesson (four groups of six students). Each answer was coded as correct, partially correct, incorrect, or blank, and each item was classified as procedural (representing, computing, or identifying) or reasoning (explaining, justifying, concluding, or stating a use).

RQ3. Descriptive statistics were computed for each test. Normality was checked with the Shapiro–Wilk test, and homogeneity of variance with Levene's test. Pre–post change in the experimental class was tested with a paired-samples t test, with the Wilcoxon signed-rank test as a robustness check. The effect size was Cohen's d for paired data, the mean difference divided by the standard deviation of the differences. Learning gain was expressed as the normalised gain [37]:

$$g = \frac{S_{post} - S_{pre}}{S_{max} - S_{pre}}$$

Here S_{pre} and S_{post} are the pre-test and post-test scores and S_{max} = 100. Gains were classified as high (g ≥ 0.7), medium (0.3 ≤ g < 0.7), or low (g < 0.3) [37]. Gain was computed for the class means and for each student.

Because post-test scores in both classes departed from normality, the two classes were compared with the Mann–Whitney U test. The effect size was $r = {Z}/{\sqrt{N}}$, supplemented by the rank-biserial correlation. Welch's t test, which does not assume equal variances, and Hedges' g were reported as a parametric check. Effect sizes were interpreted with Cohen's [38] benchmarks for small, medium, and large effects. Analyses used IBM SPSS Statistics version 29, with α = .05.

2.7 Ethical considerations

The school granted permission for the study. The study was approved by the Research Ethics Committee of the Research and Community Service (LPPM), Universitas Nahdlatul Wathan Mataram, Indonesia (approval number 058/LPPM/UNW.87/A.13/VI/2026). Written informed consent was obtained from all participating students and from their parents. Students were informed that they could withdraw at any time. The school and participants are not identified in this report, and students' faces are blurred in all photographs reproduced here.

3 Results

3.1 RQ1: AI-assisted design and conformity to the prompt

Across the four worksheets, the confirmed versions conformed to the design prompt on 6 of 48 component codes, conformed partially on 16, and deviated on 26 (Table 4). The only component that conformed in every worksheet was Let's Tell. The feelings check, Let's Guess, Let's Conclude, and the time allocation deviated in all four worksheets, and the learning objective deviated in two.

Table 4 Conformity of the confirmed worksheets to the design prompt

Prompt Component	Lesson 1	Lesson 2	Lesson 3	Lesson 4
Group identity and date	Deviates	Deviates	Partial	Partial
One-sentence objective, one competency at analysing level or above	Deviates	Deviates	Conforms	Conforms
Feelings check with three emoticons	Deviates	Deviates	Deviates	Deviates
Let's Read: illustrated three-paragraph STEAM+X story, no concept explanation	Partial	Partial	Partial	Deviates
Let's Guess: prerequisite questions	Deviates	Deviates	Deviates	Deviates
Let's Solve: open-ended contextual STEAM+X problem	Partial	Deviates	Deviates	Deviates
Two solution spaces, the second with hints	Partial	Partial	Partial	Partial
Let's Tell: presentation and peer response	Conforms	Conforms	Conforms	Conforms
Let's Conclude: correct solution, summary, infographic, formulas	Deviates	Deviates	Deviates	Deviates
Let's Practise: two open-ended problems, the second higher-order	Partial	Deviates	Partial	Deviates
Five closing reflection items	Partial	Partial	Partial	Partial
Time allocation of 15, 60, and 15 minutes	Deviates	Deviates	Deviates	Deviates
Total conforms/partial/deviates	1/5/6	1/3/8	2/5/5	2/3/7

Note. Partial = some but not all specified features present.

Three patterns underlie these codes (Fig. 2). First, the reading texts of lessons 1 and 2 used local cultural settings (a cultural festival and an arts studio in Lombok), but they explained the matrix concepts that students were meant to discover. The lesson 3 and 4 texts used school cooperative sales reports and contained no cultural element. Second, Let's Guess asked students to anticipate new content, for example what a transpose might be, rather than recall prerequisites. Third, Let's Conclude contained sentence stems for students to complete instead of the correct solution, summary, and formulas specified in the prompt.

The worksheets also contained features absent from the prompt. The lesson 1 and 2 worksheets labelled their sections with the phases epistemic prime, epistemic action, and epistemic reflection, while the lesson 3 and 4 worksheets carried no phase labels.

In the lesson 4 worksheet, all matrix notation on four slides was displayed as raw LaTeX code (for example, \begin{bmatrix} 4 & 6 \\ 8 & 10 \end{bmatrix}) rather than as rendered matrices. Smaller instances of unrendered notation also appeared in lesson 1 (a_{12} and a_{23} in Let's Practise) and lesson 2 (m \times n in Let's Conclude).

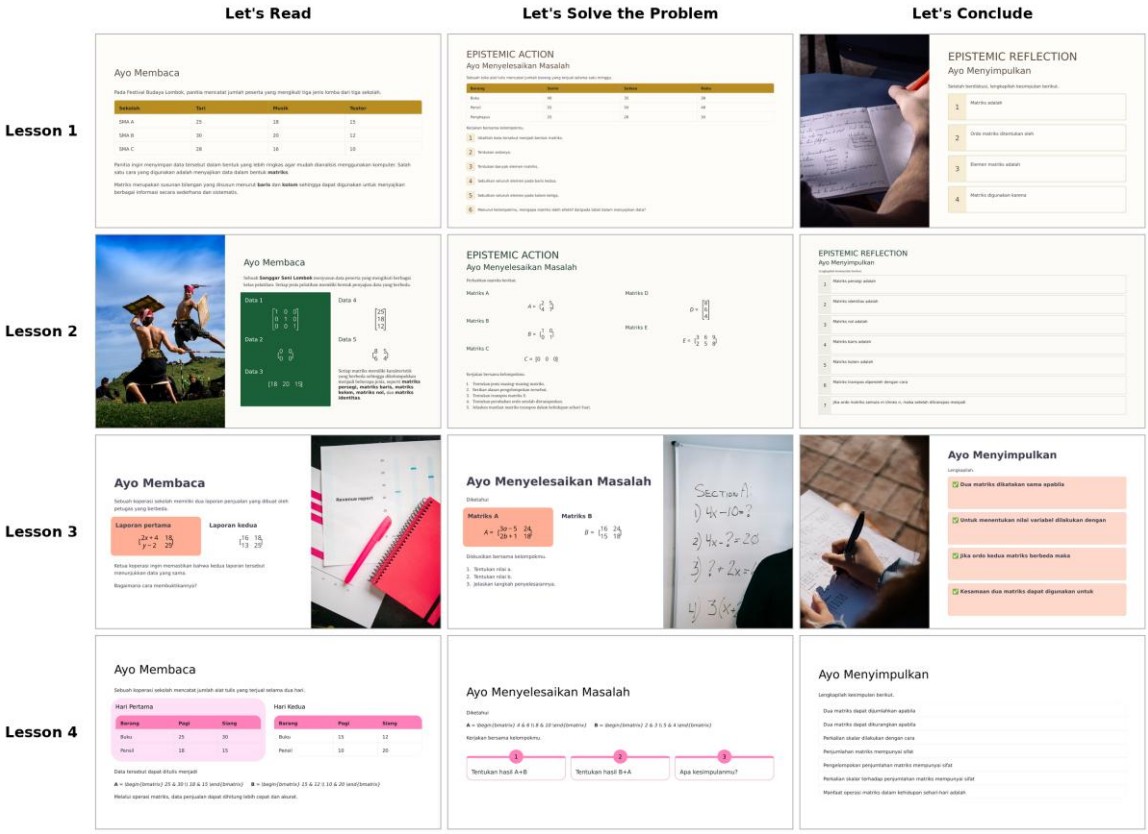

Fig. 2 Let's Read, Let's Solve the Problem, and Let's Conclude pages of the four confirmed worksheets

Figure 2 shows the shift across lessons from cultural settings in the reading texts of lessons 1 and 2 to school sales reports in lessons 3 and 4, the phase labels present only in lessons 1 and 2, the abstract matrices in the problem pages of lessons 2 to 4, the sentence stems that replaced the specified summaries in every Let's Conclude page, and the unrendered LaTeX in lesson 4.

3.2 RQ2: Functioning during classroom enactment

Reflection notes from all four lessons reported active student response. Students competed to offer opinions and to come forward to solve problems and present group

answers on the interactive board (Fig. 3). Table 5 summarises the student work recorded on the board.

Table 5 Student work recorded on the interactive board, by lesson

Lesson	Student Work Recorded on the Interactive Board
1. Concept, order, and elements	Correct 3 × 3 matrix from a sales table; correct order, number of elements, second row, and third column. Reasoning question left blank. In Let's Conclude, matrix and order defined correctly; element defined as "the number of data"; final stem left blank
2. Types and transpose	Not legible in the photographs
3. Equality of two matrices	As far as legible, linear equations derived from equal matrices and solved for the unknowns
4. Algebraic operations	(A + B) + C and A + (B + C) computed correctly for teacher-supplied matrices; associative property stated

Two observations qualify the written record in Table 5. First, the board did not show an answer to the question of why a matrix presents data more effectively than a table; that group answered it orally during its presentation. Second, the lesson 1 definition of an element on the board conflated an element with the number of elements. In lesson 4, the teacher replaced the unreadable matrices with A = [3 2; 5 1], B = [4 −1; 2 3], and C = [0 1; 3 2], and students used these to compute both sides of the associative property.

The reflections also recorded that some students remained passive. The main concern raised across the reflections was keeping lessons within the planned time, and the team agreed to use a stopwatch to keep each phase within its allocation.



Fig. 3 Students working at the interactive board

The completed worksheets give a fuller written record than the board. Across the four group worksheets, the groups answered 103 items: 84 correctly, 8 partially, and 2 incorrectly, and left 9 blank (Table 6). Procedural items were almost always correct (67 of 75). Reasoning items were answered correctly less often (17 of 28), and lesson 4 accounted for 8 of the 9 blank items.

Table 6 Coding of answers in the completed group worksheets

Worksheet or Item Type	Items	Correct	Partial	Incorrect	Blank
Lesson 1	22	22	0	0	0
Lesson 2	25	19	3	2	1
Lesson 3	22	22	0	0	0
Lesson 4	34	21	5	0	8
Procedural items	75	67	2	1	5
Reasoning items	28	17	6	1	4
Total	103	84	8	2	9

Written reasoning was present but varied in depth. In lesson 1, the group justified matrices as making data “easier to calculate and analyse” (*gampang menghitung dan menganalisis*) and defined an element correctly as “the values in the matrix” (*nilai yang ada di dalam matriks*), unlike the group whose board work is described in Table 5. The lesson 3 group gave the fullest justifications, explaining that equal orders ensure that “each element has a matching position pair” (*setiap elemen memiliki pasangan posisi*

yang pas), and wrote a paragraph-length alternative solution; it also solved the Let's Read equations without being asked. The lesson 2 group classified the square matrix A = [2 5; 4 7] as a row matrix, carried this error into its answers on order changes, did not identify the column matrix in Let's Guess, and described its new knowledge as "inverse matrices" (*invers matriks*), although only the transpose had been taught. In lesson 4, the group stated commutativity informally as "basically it can be reversed" (*intinya dia bisa dibalik*) and associativity as "the method is the same" (*metodenya sama*). It left Let's Guess, the scalar-multiplication pattern, the alternative solution, and the explanation items in Let's Practise blank, and marked three proofs as done without showing working. The same group rewrote every coded matrix in standard notation before computing with it (Fig. 4).

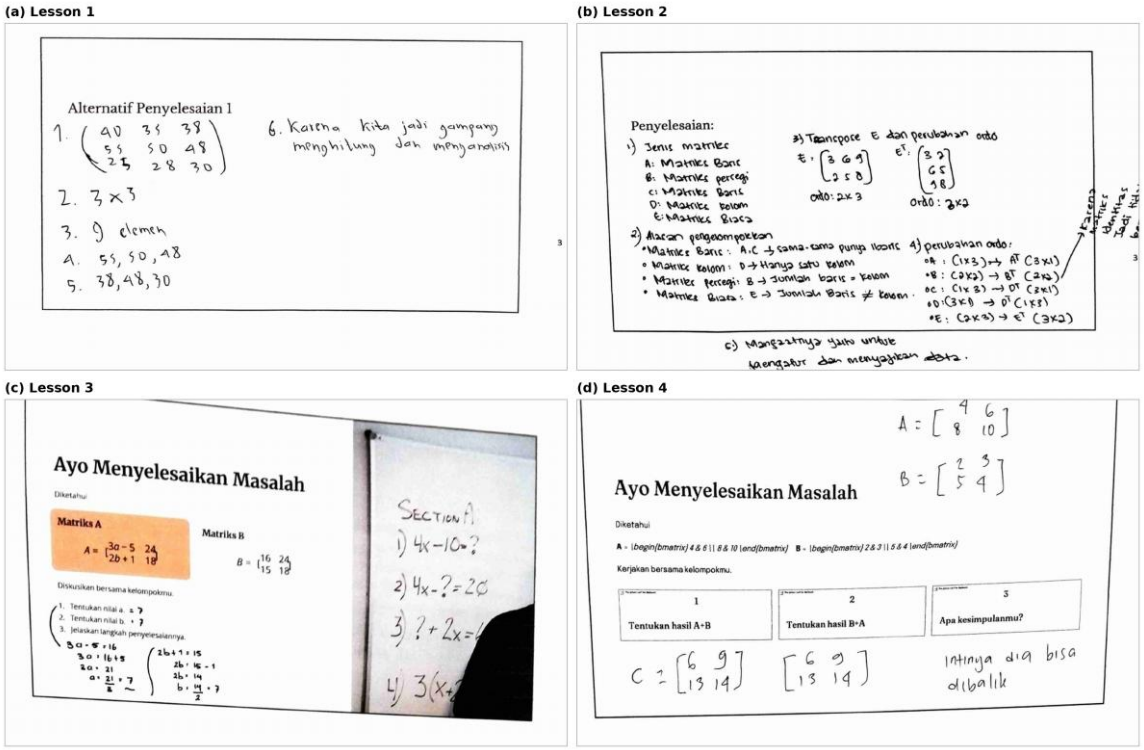

Fig. 4 Excerpts from completed group worksheets

Figure 4 shows that written work was fullest in lesson 3 and sparsest in lesson 4, and that the lesson 4 group wrote the coded matrices out by hand before working with them.

3.3 RQ3: Achievement

In the experimental class, mean scores rose from the pre-test to the post-test, and every one of the 36 students scored higher on the post-test (Table 7). Pre-test and post-test scores were not normally distributed, but the pre–post differences were (Shapiro–Wilk W = 0.956, p = .157).

Table 7 Descriptive statistics

Test	n	M	SD	Mdn	Min–Max	Shapiro–Wilk p
Experimental pre-test	36	34.03	21.94	27.5	5–90	.015
Experimental post-test	36	70.28	20.21	75.0	25–100	.023
Control post-test	35	44.00	24.79	35.0	10–85	.002

The pre–post difference was statistically significant, t(35) = 11.57, p < .001, with a mean gain of 36.25 points (95% CI [29.89, 42.61]) and a paired effect size of d = 1.93. The Wilcoxon signed-rank test agreed, Z = 5.24, p < .001, with no negative ranks. The class normalised gain was 0.55, in the medium category. Individual gains averaged 0.58 (SD = 0.26, Mdn = 0.64): 14 students (38.9%) reached a high gain, 15 (41.7%) a medium gain, and 7 (19.4%) a low gain. Pre-test and post-test scores correlated at r = .60 (p < .001).

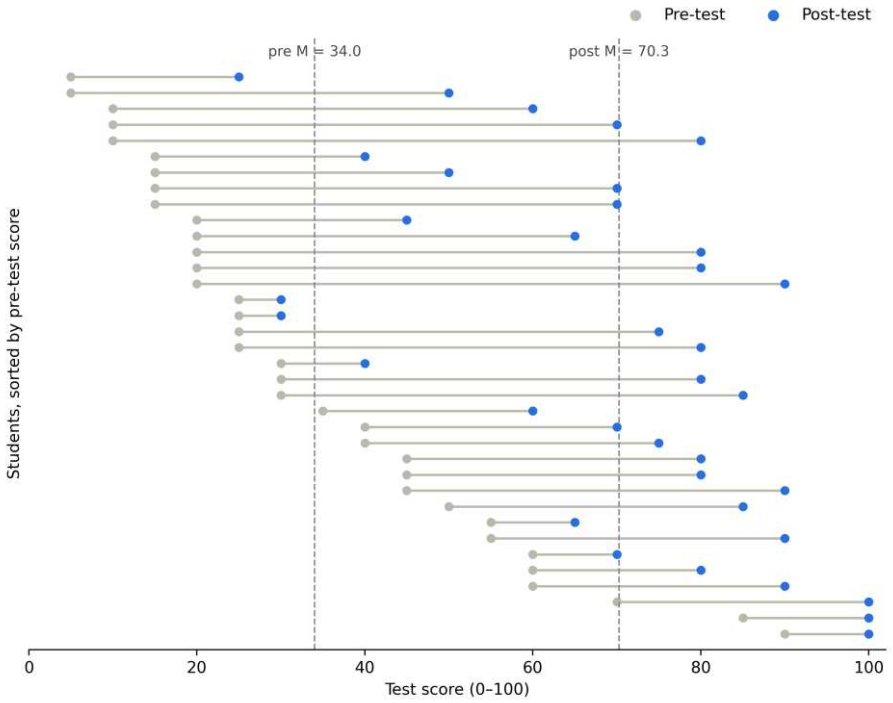

Fig. 5 Pre-test and post-test scores of each student in the experimental class

Figure 5 shows that the gains were spread across the whole class rather than concentrated among students with low pre-test scores. The smallest gains, of 5 to 10 points, occurred at both low and middle pre-test levels.

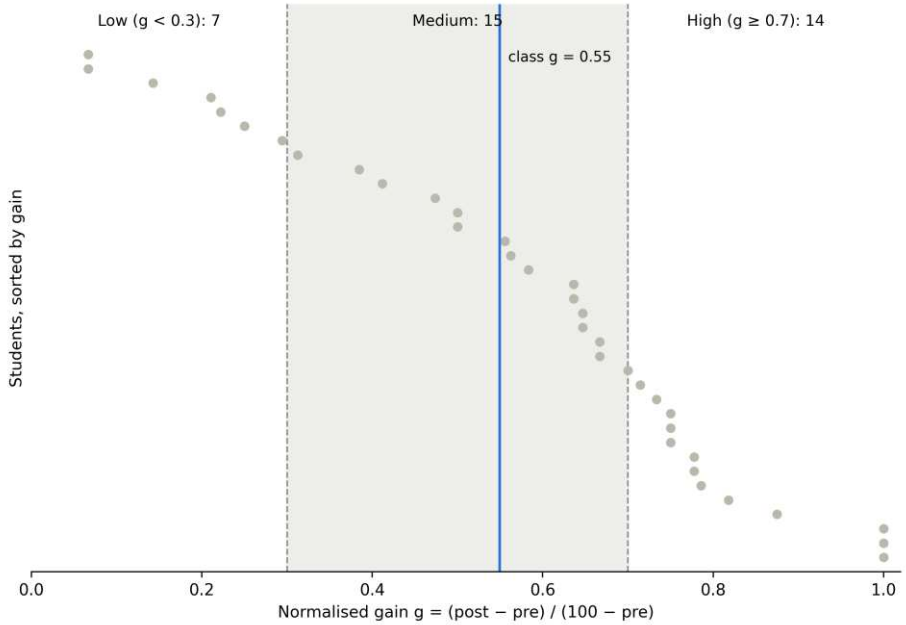

Fig. 6 Distribution of individual normalised gains in the experimental class

Figure 6 shows that individual gains were spread across all three categories, with three students reaching the maximum gain of 1.0 and seven students remaining in the low category.

On the post-test, the experimental class scored higher than the control class (Table 8). Because the control class took no pre-test, this comparison concerns post-test levels only, not gains.

Table 8 Post-test comparison between classes

Test	Statistic	p	Effect size
Mann–Whitney U	U = 271, Z = 4.14	< .001	r = .49; rank-biserial = .57
Welch's t	t(65.55) = 4.89	< .001	Hedges' g = 1.15
Levene (median-centred)	F(1, 69) = 2.82	.098	—
Levene (mean-centred)	F(1, 69) = 4.53	.037	—

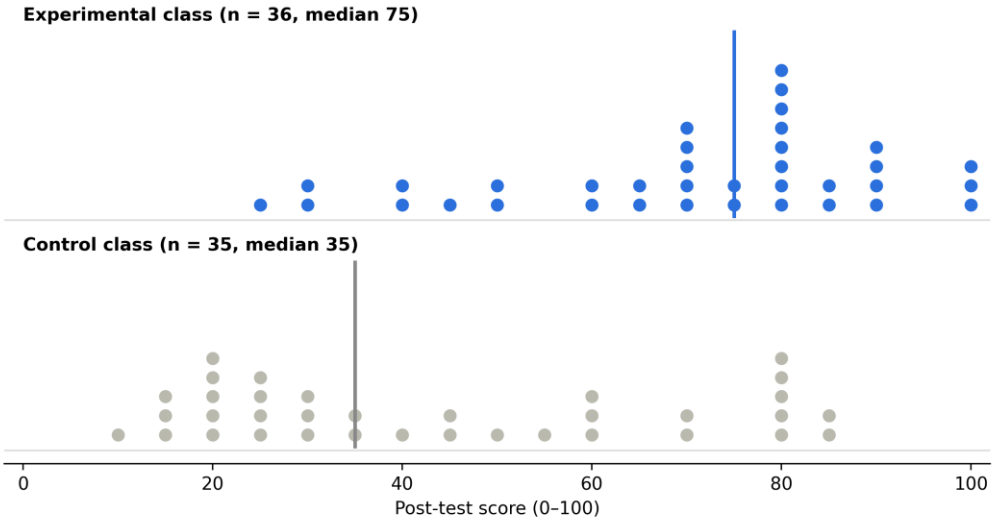

Fig. 7 Post-test scores of the experimental and control classes

Figure 7 shows that the two distributions overlapped. A cluster of control students scored 70 or higher, while several experimental students scored below 50.

4 Discussion

The central finding of this study is a contrast between design fidelity and classroom functioning. The confirmed worksheets followed the structured prompt on only a small share of its specifications, yet the lessons built on them produced sustained student participation, and every student in the experimental class scored higher after the four lessons. What students experienced therefore depended less on how faithfully the AI realised the prompt than on what teachers did with its output, before and during the lessons. This is the sense in which AI drafted and teachers decided.

The findings also bear on one another in ways that are not visible when each is read alone. The component that deviated most consistently, Let's Conclude, replaced the specified correct summary with sentence stems for students to complete. This placed the formulation of concepts in students' hands, which is consistent with the epistemic intent of the design, but it also removed the written reference against which a student's formulation could be checked. The misconception recorded on the board in lesson 1, where an element was defined as the number of data items, arose at exactly this point, as did the lesson 2 group's description of the transpose as an inverse. A second link

concerns reasoning. The board alone suggested that justification happened mainly in talk, but the completed worksheets show that written reasoning was more common: 17 of 28 reasoning items were answered correctly, although usually in a short phrase. Written reasoning was fullest in lesson 3, whose prompts asked directly for conditions and consequences, and sparsest in lesson 4, where students first had to decode the notation and the reasoning prompts came last in a long worksheet. Because the post-test asked for written explanations in every item, item-level scores, which were not available, would be needed to judge how far this reasoning carried over to the test. A third link concerns participation. Reflection notes described students competing to present, yet some students remained passive, and seven students had low normalised gains. These two observations may describe different students. Visible enthusiasm may have been concentrated among those already able to contribute, although the data do not allow individual matching.

These patterns become clearer when set against previous research. Studies of ChatGPT in mathematics lesson planning report that its output requires substantial revision and that teachers' expertise remains decisive [4, 13], consistent with research showing that teachers act as designers who adapt the materials they are given [17]. The present study supports this conclusion but locates part of the needed revision later than planning studies can observe. The unrendered LaTeX in the lesson 4 worksheet was not a mathematical error of the kind documented by Pardos and Bhandari [15]. It was a formatting failure that passed through teacher and researcher confirmation and surfaced only on the classroom display, where the teacher repaired it by substituting new matrices and one group rewrote the coded matrices in standard notation before computing with them. Cheng [35] found that ChatGPT made collaborative planning within lesson study more efficient and was followed by improved concept test results. Our findings are consistent with that report and extend it in two ways: to mathematics worksheets rather than lesson plans, and to the moment of enactment as a site where teacher judgement corrected AI output. Isnawan et al. [30] likewise reported that an epistemic learning sequence developed through lesson study produced active student contributions but ran short of time, so that the practice activity had to be dropped. The time pressure recorded here suggests that this may be a feature of sequences that give students extended presentation and response time, not of a single implementation. Meta-analytic evidence points to positive effects of

ChatGPT-supported instruction on academic performance [16], but such syntheses cannot show what teachers changed along the way; the present study suggests that these changes may be part of what produces the effect.

Several explanations for these patterns are plausible, although each remains tentative. The high level of participation may reflect the structure of Let's Tell, which made presenting and responding a routine part of every lesson, combined with the interactive board, which made students' work public and easy to revise, in line with the way digital technologies are redefining learning spaces in mathematics classrooms [39]. Active learning strategies in which students present and discuss their own work have been linked to similar engagement in mathematics classrooms [40], and lesson study helped teachers orchestrate such whole-class discussions [32]. Novelty may also have contributed, and four lessons are too few to separate these influences. The fading of the cultural element across lessons has a more specific explanation. Local culture appeared in the reading texts of lessons 1 and 2, which dealt with representing data, but not in lessons 3 and 4, which dealt with equality of matrices and algebraic properties. Abstract operations are harder to tie to a cultural setting, and without a specific cultural anchor in the prompt the confirmed worksheets fell back on generic school and workplace contexts. This echoes evidence that disciplines in integrated STEM curricula are often only loosely or contextually integrated in practice [24]. It also contrasts with ethnomathematics work in which the mathematics is drawn from the cultural artefact itself, rather than placed beside it [25, 27].

The achievement results should be read with equal care. The pre–post gain and the post-test advantage over the control class were large, but three features of the evidence limit what they can show. The control class took no pre-test, so its starting level is known only from the teacher's report. Some test content overlapped with worksheet activities, including pre-test item 4, which used the same matrices and context as the lesson 4 worksheet. Three students also reached the maximum score, so the test could not register further growth at the top. The gains are therefore best treated as supporting evidence that the design was compatible with learning, not as an estimate of its effect.

In theoretical terms, the study supports the ELP sequence at the level of design rather than testing it. Read through the theory of didactical situations on which ELP builds [30], the lessons showed strong situations of action and formulation, as students solved problems and explained their solutions, and a working situation of validation, as peers responded during Let's Tell. Institutionalisation, in which the teacher establishes the shared, correct form of knowledge, was the weakest phase, and the misconceptions in lessons 1 and 2 show the cost of leaving it implicit. Ethnomathematics-based designs similarly treat institutionalisation as the decisive step in turning cultural activity into school concepts [41]. With respect to apt epistemic performance [29], the clearest evidence concerns participating in knowledge construction with others. Reliable reasoning was clearest in procedural work, where 67 of 75 worksheet items were correct, and less consistent in written justification, and metacognitive regulation was not measured.

Several implications follow. For teachers who start from AI drafts, the findings suggest checking every worksheet on the device on which it will be shown, and keeping an explicit moment in which the teacher confirms the correct concept after students have formulated their own. For designers of prompts, the deviations in Table 4 indicate which specifications AI tools tend to drop: prerequisite-focused questions, culturally anchored problems, correct summaries, and time allocations. Prompts may need to state these as non-negotiable constraints, name the cultural element for each lesson, and specify an output format that displays mathematical notation reliably. For researchers, the loss of the raw AI output in this study shows the value of retaining every version of an AI-assisted material, so that the contributions of the AI and of teachers can be separated.

The findings should, however, be interpreted in light of several limitations. The study involved one school, one teacher, and one macro-cycle of design research, and its findings may not transfer to other settings. The raw AI outputs were not retained, so the conformity analysis describes the confirmed worksheets and cannot separate AI deviations from teacher revisions. The photographs of lesson 2 were not legible, and completed worksheets were available for only one group per lesson. The control class had no pre-test, and equivalence rested on the teacher's report. On the instruments, test content overlapped with worksheet activities, sub-question 2d was not parallel

across forms, a single teacher scored all answers without a reliability check, and the printed post-test carried a header naming the wrong topic and an instruction referring to the pre-test. Agreement between the two coders before discussion was not quantified. Future research should complete a second macro-cycle with a revised prompt and retain all versions of the materials. It should also pre-test all classes, or use random assignment where possible, use tests that do not reuse worksheet content, and analyse written justification at item level with more than one scorer. Further work could also examine directly which students remain passive in highly participatory lessons.

5 Conclusion

Generative AI produced complete ELP-STEAM+X worksheets on matrices from a structured prompt, but the versions confirmed by teachers and researchers realised only a small share of the prompt's specifications. The collaborative presentation task survived intact, whereas prerequisite questions, culturally anchored problems, correct concept summaries, and time allocations were most often lost. In classroom enactment across four lesson study cycles, the worksheets supported active participation and largely correct procedural work. Written justification, however, was often brief and largely missing in the final lesson, misconceptions appeared in two lessons, some students stayed passive, and unrendered notation had to be repaired by the teacher and by students. Every student in the experimental class scored higher after the four lessons, with a medium normalised gain, and the class outperformed a conventionally taught class on the post-test.

The main contribution of this study lies in tracing AI-drafted materials from prompt to classroom. Doing so shows that their value depended on teachers' decisions at two points: when the drafts were confirmed and when they met students, particularly at the moment of establishing the correct concept. Because the study involved one school and one design cycle, gave the comparison class no pre-test, and used tests that overlapped with worksheet content, the achievement results indicate that the design was compatible with learning, not how large its effect was. A second design cycle should use a prompt that states its key specifications as firm constraints, retain every version of the materials, pre-test all classes, and analyse students' written reasoning item by item.

Acknowledgements

The authors thank the model teacher, the observer teacher, the participating students, and the 25 teachers who took part in the focus group discussion. The authors also thank Universitas Nahdlatul Wathan Mataram and the University of Jeddah for their support.

Author contributions

Muhamad Galang Isnawan: conceptualisation, methodology, design prompt development, investigation, formal analysis, visualisation, and writing (original draft). Naif Mastoor Alsulami: methodology, independent coding of the worksheets, validation of the analysis, and writing (review and editing). Lalu Sirajul Hadi: investigation, formal analysis, visualisation, and writing (review and editing). Sudirman: methodology and writing (review and editing). Samsuriadi: investigation, visualisation, and writing (original draft). R. Didi Kuswara: conceptualisation, methodology, design prompt development, and investigation. I Ketut Sukarma: conceptualisation, methodology, design prompt development, and writing (original draft). All authors read and approved the final manuscript.

Funding

This research and its publication were supported by Universitas Nahdlatul Wathan Mataram, Indonesia; the University of Jeddah, Saudi Arabia; Universitas Mataram, Indonesia; and Universitas Pendidikan Mandalika, Indonesia.

Data availability

The test scores and the design prompt analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study was approved by the Research Ethics Committee of the Institute for Research and Community Service (LPPM), Universitas Nahdlatul Wathan Mataram, Indonesia (approval number 058/LPPM/UNW.87/A.13/VI/2026), and the school granted permission for the study. Written informed consent was obtained from all participating students and from their parents, and students were informed that they could withdraw at any time. All procedures were performed in accordance with relevant guidelines and regulations. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki.

Consent for publication

Not applicable. The article contains no identifiable personal data; the school and participants are not named, and students' faces are blurred in all photographs.

Competing interests

The authors declare no competing interests.

References

1. Fütterer T, Fischer C, Alekseeva A, Chen X, Tate T, Warschauer M, et al. ChatGPT in education: global reactions to AI innovations. Scientific Reports. 2023;13:15310. https://doi.org/10.1038/s41598-023-42227-6
2. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences. 2023;103:102274. https://doi.org/10.1016/j.lindif.2023.102274
3. Farrokhnia M, Banihashem SK, Noroozi O, Wals A. A SWOT analysis of ChatGPT: implications for educational practice and research. Innovations in

Education and Teaching International. 2024;61(3):460–474. https://doi.org/10.1080/14703297.2023.2195846

4. Pepin B, Buchholtz N, Salinas-Hernández U. A scoping survey of ChatGPT in mathematics education. Digital Experiences in Mathematics Education. 2025;11:9–41. https://doi.org/10.1007/s40751-025-00172-1
5. Bond M, Khosravi H, De Laat M, Bergdahl N, Negrea V, Oxley E, et al. A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. International Journal of Educational Technology in Higher Education. 2024;21:4. https://doi.org/10.1186/s41239-023-00436-z
6. Celik I, Dindar M, Muukkonen H, Järvelä S. The promises and challenges of artificial intelligence for teachers: a systematic review of research. TechTrends. 2022;66(4):616–630. https://doi.org/10.1007/s11528-022-00715-y
7. Heung YME, Chiu TKF. How ChatGPT impacts student engagement from a systematic review and meta-analysis study. Computers and Education: Artificial Intelligence. 2025;8:100361. https://doi.org/10.1016/j.caeai.2025.100361
8. Lo CK, Hew KF, Jong MSY. The influence of ChatGPT on student engagement: a systematic review and future research agenda. Computers & Education. 2024;219:105100. https://doi.org/10.1016/j.compedu.2024.105100
9. Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences. 2023;13(4):410. https://doi.org/10.3390/educsci13040410
10. Wardat Y, Tashtoush MA, AlAli R, Jarrah AM. ChatGPT: a revolutionary tool for teaching and learning mathematics. Eurasia Journal of Mathematics, Science and Technology Education. 2023;19(7):em2286. https://doi.org/10.29333/ejmste/13272
11. Almarashdi HS, Jarrah AM, Abu Khurma O, Gningue SM. Unveiling the potential: a systematic review of ChatGPT in transforming mathematics teaching and learning. Eurasia Journal of Mathematics, Science and Technology Education. 2024;20(12):em2555. https://doi.org/10.29333/ejmste/15739
12. Opesemowo OAG, Adewuyi HO. A systematic review of artificial intelligence in mathematics education: the emergence of 4IR. Eurasia Journal of Mathematics, Science and Technology Education. 2024;20(7):em2478. https://doi.org/10.29333/ejmste/14762

13. Gurl TJ, Markinson MP, Artzt AF. Using ChatGPT as a lesson planning assistant with preservice secondary mathematics teachers. Digital Experiences in Mathematics Education. 2025;11:114–139. https://doi.org/10.1007/s40751-024-00162-9
14. van den Berg G, du Plessis E. ChatGPT and generative AI: possibilities for its contribution to lesson planning, critical thinking and openness in teacher education. Education Sciences. 2023;13(10):998. https://doi.org/10.3390/educsci13100998
15. Pardos ZA, Bhandari S. ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. PLOS ONE. 2024;19(5):e0304013. https://doi.org/10.1371/journal.pone.0304013
16. Deng R, Jiang M, Yu X, Lu Y, Liu S. Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. Computers & Education. 2025;227:105224. https://doi.org/10.1016/j.compedu.2024.105224
17. Jukić Matić L. Navigating the seas of the curriculum: a mathematics teacher's interactions with curriculum materials. European Journal of Science and Mathematics Education. 2024;12(4):468–488. https://doi.org/10.30935/scimath/14912
18. Mishra P, Warr M, Islam R. TPACK in the age of ChatGPT and generative AI. Journal of Digital Learning in Teacher Education. 2023;39(4):235–251. https://doi.org/10.1080/21532974.2023.2247480
19. Lan YJ, Chen NS. Teachers' agency in the era of LLM and generative AI: designing pedagogical AI agents. Educational Technology & Society. 2024;27(1):I–XVIII. https://doi.org/10.30191/ETS.202401_27(1).PP01
20. Ellis AB, Özgür Z. Trends, insights, and developments in research on the teaching and learning of algebra. ZDM Mathematics Education. 2024;56:199–210. https://doi.org/10.1007/s11858-023-01545-9
21. Marín-Marín JA, Moreno-Guerrero AJ, Dúo-Terrón P, López-Belmonte J. STEAM in education: a bibliometric analysis of performance and co-words in Web of Science. International Journal of STEM Education. 2021;8:41. https://doi.org/10.1186/s40594-021-00296-x
22. Belbase S, Mainali BR, Kasemsukpipat W, Tairab H, Gochoo M, Jarrah A. At the dawn of science, technology, engineering, arts, and mathematics (STEAM) education: prospects, priorities, processes, and problems. International Journal

of Mathematical Education in Science and Technology. 2022;53(11):2919–2955. https://doi.org/10.1080/0020739X.2021.1922943

23. Martín-Cudero D, Cid-Cid AI, Guede-Cid R. Analysis of mathematics education from a STEAM approach at secondary and pre-universitary educational levels: a systematic review. Journal of Technology and Science Education. 2024;14(2):507–528. https://doi.org/10.3926/jotse.2349
24. Roehrig GH, Dare EA, Ring-Whalen E, Wieselmann JR. Understanding coherence and integration in integrated STEM curriculum. International Journal of STEM Education. 2021;8:2. https://doi.org/10.1186/s40594-020-00259-8
25. Prahmana RCI, Istiandaru A. Learning sets theory using shadow puppet: a study of Javanese ethnomathematics. Mathematics. 2021;9(22):2938. https://doi.org/10.3390/math9222938
26. Umbara U, Wahyudin W, Prabawanto S. Exploring ethnomathematics with ethnomodeling methodological approach: how does Cigugur indigenous people using calculations to determine good day to build houses. Eurasia Journal of Mathematics, Science and Technology Education. 2021;17(2):em1939. https://doi.org/10.29333/ejmste/9673
27. Prayitno S, Hayati L, Novitasari D, Lu'luilmaknun U. Wariga ethnomathematics: mathematical structures and Sasak cultural values in the traditional calendar system. Foro de Educación. 2026;24(1):320–340. https://doi.org/10.14201/fde.24116
28. El Bedewy S, Lavicza Z. STEAM + X – Extending the transdisciplinary of STEAM-based educational approaches: a theoretical contribution. Thinking Skills and Creativity. 2023;48:101299. https://doi.org/10.1016/j.tsc.2023.101299
29. Barzilai S, Chinn CA. The AIR and Apt-AIR frameworks of epistemic performance and growth: reflections on educational theory development. Educational Psychology Review. 2024;36:91. https://doi.org/10.1007/s10648-024-09927-5
30. Isnawan MG, Alsulami NM, Rasilah, Sukarma IK, Lavicza Z. Didactic design research through lesson study activities: STEM-based courses for representative abilities of prospective mathematics teachers. European Journal of STEM Education. 2025;10(1):12. https://doi.org/10.20897/ejsteme/16758
31. Ding M, Huang R, Pressimone Beckowski C, Li X, Li Y. A review of lesson study in mathematics education from 2015 to 2022: implementation and impact. ZDM

Mathematics Education. 2024;56(1):87–99. https://doi.org/10.1007/s11858-023-01538-8

32. Faria FAB, da Ponte JP, Rodrigues M. Teachers' leading whole-class discussions in a mathematics lesson study: from initial understanding to orchestration in practice. European Journal of Science and Mathematics Education. 2024;12(1):156–174. https://doi.org/10.30935/scimath/14149
33. Hoadley C, Campos FC. Design-based research: what it is and why it matters to studying online learning. Educational Psychologist. 2022;57(3):207–220. https://doi.org/10.1080/00461520.2022.2079128
34. McKenney S, Reeves TC. Educational design research: portraying, conducting, and enhancing productive scholarship. Medical Education. 2021;55(1):82–92. https://doi.org/10.1111/medu.14280
35. Cheng ECK. Enhancing teachers' TPACK through ChatGPT-supported lesson study. International Journal for Lesson and Learning Studies. 2025;14(3):215–236. https://doi.org/10.1108/IJLLS-11-2024-0277
36. Şimşek N. Integration of ChatGPT in mathematical story-focused 5E lesson planning: teachers and pre-service teachers' interactions with ChatGPT. Education and Information Technologies. 2025;30(8):11391–11462. https://doi.org/10.1007/s10639-024-13258-x
37. Hake RR. Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics. 1998;66(1):64–74. https://doi.org/10.1119/1.18809
38. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates; 1988.
39. Engelbrecht J, Borba MC. Recent developments in using digital technology in mathematics education. ZDM Mathematics Education. 2024;56:281–292. https://doi.org/10.1007/s11858-023-01530-2
40. Vale I, Barbosa A. Active learning strategies for an effective mathematics teaching and learning. European Journal of Science and Mathematics Education. 2023;11(3):573–588. https://doi.org/10.30935/scimath/13135
41. Sudirman, Rodríguez-Nieto CA, Bonyah E. Integrating ethnomathematics and ethnomodeling in institutionalization of school mathematics concepts: a study of fishermen community activities. Journal on Mathematics Education. 2024;15(3):835–858. https://doi.org/10.22342/jme.v15i3.pp835-858