Exploring the impact of ChatGPT on college students' mathematical creativity: A quasi-experimental study

Zhiwei Liu, Gang Li, Haode Zuo, Yongjing Lu
College of Mathematical Science, Yangzhou University, Yangzhou, China

Jue Feng
College of Educational Science, Yangzhou University, Yangzhou, China

Recent advancements in the use of ChatGPT in mathematics education have drawn significant attention from educators and researchers. However, the effect of ChatGPT on college students' mathematical creativity remains inconclusive. Based on the hybrid human-AI regulation theory, this study addressed this gap through a quasi-experiment comparing an experimental group using ChatGPT to solve a mathematical creativity task with a control group that did not. The results showed that ChatGPT had a positive effect on college students' mathematical creativity and enhanced their self-efficacy and interest in task resolution. Also, students using ChatGPT found the task easier and required less mental effort. Notably, ChatGPT may hurt the accuracy of metacognitive evaluation. Overall, the findings revealed that ChatGPT positively influences mathematical creativity, but it requires greater use of metacognitive skills. This study further refined the hybrid human-AI regulation theory by showing that improved performance and inaccurate metacognitive evaluation may coexist in human–ChatGPT co-creation. The findings offer valuable insights for guidance on using ChatGPT to enhance college students' mathematical creativity.

Implications for practice or policy:

- Mathematics educators could use ChatGPT to enhance college students' mathematical creativity, but this may impair their metacognitive evaluation.
- Policymakers should integrate ChatGPT into mathematics curriculum and develop courses in artificial intelligence literacy to enhance students' metacognitive skills in the human–ChatGPT co-creation process.
- Educators could adopt a "humans first, ChatGPT second" approach to mathematics instruction, encouraging students to actively engage in metacognitive regulation during the human–ChatGPT co-creation process.

Keywords: human–ChatGPT co-creation, mathematical creativity, mathematical problem solving, metacognitive skills, mathematics education

Introduction

In mathematics education, mathematical creativity tasks (MCT) are widely used to foster students' mathematical creativity (Molad et al., 2020). These tasks often lack fixed solutions and require students to engage in flexible reasoning and creative thinking within complex mathematical contexts (Levav-Waynberg & Leikin, 2012). With the advancement of educational theories and emerging technologies, mathematics education has increasingly focused on enhancing students' creative thinking through MCT (Leikin & Sriraman, 2022). As an emerging technology, ChatGPT can assist students in solving mathematical tasks by providing useful information and generating prototypical solutions (Zhou et al., 2025).

However, research on the impact of ChatGPT on college students' performance in solving MCT remains scarce. Also, no studies to date have examined the mechanisms by which metacognitive regulation strategies influence the human–ChatGPT co-creation process. Research has shown that solving MCT relies not only on students' creative thinking but also on their metacognitive skills (Bahar et al., 2024). It requires students to use divergent thinking to propose multiple prototypical solutions and employ convergent thinking to select the optimal one (Guilford, 1967; Habib et al., 2024). Students should continuously monitor their problem-solving progress, adjust their cognitive strategies and engage in problem reconstruction and self-evaluation as the task context evolves (Winne, 2017). In other words, solving MCT requires a combination of creative thinking and metacognitive skills.

Therefore, this study aimed to explore the impact of ChatGPT on college students' performance in solving MCT and to uncover how metacognitive regulation strategies shape this process. More importantly, drawing from the hybrid human-AI regulation theory (Molenaar, 2022a, 2022b), this study also investigated five key factors that are required for efficient use of ChatGPT: (a) self-efficacy in task resolution, (b) accuracy of self-evaluation, (c) perceived task interest, (d) perceived task difficulty and (e) invested mental effort. Therefore, this study explored the following research questions (RQs):

- RQ1: Does ChatGPT influence college students' mathematical creativity?
- RQ2: Does ChatGPT affect college students' self-efficacy in solving MCT?
- RQ3: Does ChatGPT influence college students' evaluation (e.g., accuracy of self-evaluation, task interest and difficulty, and invested mental effort) of their co-creation processes?

Literature review

ChatGPT

ChatGPT is a generative artifical intelligence (AI) of large language models, and its core version includes GPT-4o (OpenAI, 2023). Its massive number of parameters enables it to generate responses that closely align with user expectations (Hitsuwari et al., 2023). However, this also introduces risks, as it may produce outcomes that appear plausible but are inaccurate (Liu et al., 2025). There are two reasons for this. First, the extensive sources of ChatGPT's training data may be controversial. Second, ChatGPT is trained using reinforcement learning from human feedback, in which evaluators reward or penalise ChatGPT based on goal achievement. Because goals may conflict, ChatGPT often generates responses that are partially true yet seemingly plausible to meet user expectations, but not necessarily completely accurate (Fernandes et al., 2026). This phenomenon is especially evident in speculative or hypothetical inquiries. When prompts fall outside the scope of its training data, ChatGPT may fabricate coherent but unfounded information. Therefore, ChatGPT may carry significant risks in education, as students may be unable to identify inaccurate information, and relevant research has issued warnings about this (Peters et al., 2024). For example, Niloy et al. (2024) revealed a detrimental association between ChatGPT use and college students' creative writing abilities. This observation was confirmed by a study that found that human-crowdsourced solutions outperformed ChatGPT in novelty (Boussioux et al., 2024). The decline in novelty when co-creating with ChatGPT has also been demonstrated in the arts (Zhou & Lee, 2024).

However, ChatGPT also has many advantages. Habib et al. (2024) observed that students using ChatGPT achieved higher scores on creative tests. Urban et al. (2024) also found that students using ChatGPT produced more original solutions in creative problem-solving tasks. Similarly, Noy and Zhang (2023) reported that ChatGPT enhanced students' self-efficacy in solving creative tasks, resulting in their spending less time on the tasks. Nevertheless, the fact that ChatGPT may generate partially accurate answers to appear credible and avoid penalties, makes it necessary for students to actively monitor and evaluate the information they obtain from ChatGPT (Bezirhan & von Davier, 2023). According to the hybrid human-AI regulation theory (Molenaar, 2022a, 2022b), students should assess whether each part of the response fits within the context of the problem-solving task. This requires metacognitive regulation, including processes such as selecting relevant information, rephrasing and clarifying ideas, debugging errors and adjusting strategies to ensure coherence and meaning. Finally, students need to perform accurate self-evaluation to judge the originality of outcomes. In other words, with the emergence of ChatGPT, metacognition may become increasingly important in solving creative tasks.

Mathematical creativity and open-ended tasks

Guilford's (1967) seminal structure of intellect model identified three core components of creativity: fluency, flexibility and originality. This framework has significantly affected research in mathematics education, where the same components are used to characterise and evaluate students' mathematical creativity by engaging them with mathematical open-ended tasks (Leikin & Sriraman, 2022; Ron-Ezra & Levenson, 2025). A mathematical open-ended task is designed to elicit several distinct approaches or representations for the same problem and to operationalise Guilford's framework by valuing the quantity and variety of mathematical ideas produced, thereby fostering an environment that encourages creative exploration (Leikin & Elgrably, 2022). It is also designed to be vague and open, with minimal data, allowing for multiple solutions, improvisation and flexible thinking (Suherman & Vidákovich, 2022). Students' responses in such tasks are typically assessed based on the three indicators of creativity proposed by Kattou et al. (2013). Therefore, this study measured college students' mathematical creativity by the fluency, flexibility and originality of the solutions they produced in mathematical open-ended tasks (García-García et al., 2024).

The effects of metacognition on solving MCT

In solving MCT, students employ metacognitive monitoring when setting unique goals ("What should my ideas look like?") and developing original problem-solving strategies ("What should I do to create the solution I want?") (Greene et al., 2019). Then, they apply metacognitive regulation to various prototypical solutions they generate, selecting the most promising ones for further refinement. Finally, they self-evaluate their outcomes ("Is this the outcome I planned at the beginning?"), making accurate self-evaluation a crucial component of highly creative problem-solving performance (Urban & Urban, 2023b)

Various metacognitive experiences (e.g., perceived task interest and difficulty) provide valuable cues for students using ChatGPT, helping to guide the allocation and regulation of cognitive resources (e.g., invested mental effort) during the problem-solving process (Puente-Díaz et al., 2021). Perceived task difficulty refers to students' subjective evaluation of the difficulty of a specific mathematical task. It may be influenced by personal factors (e.g., self-efficacy, prior experience) (Winne, 2017). Perceived task interest refers to students' belief that the task is engaging or capable of arousing curiosity. When students find a task interesting, they are more likely to invest greater mental effort (Hawlitschek & Joeckel, 2017). Mental effort is defined as the cognitive energy and resources students invest in completing a task. If the task is neither too simple nor too complex, there is a strong correlation between invested mental effort and perceived task difficulty (Scheiter et al., 2020).

Additionally, self-efficacy refers to students' confidence in solving MCT (Bicer et al., 2020). Higher self-efficacy helps individuals overcome challenges in solving complex problems (Schunk & Mullen, 2012). It is worth noting that when self-efficacy is excessively high, a lack of self-doubt regarding an individual's skills may undermine motivation, negatively affecting their performance (Suherman & Vidákovich, 2024).

The present study

Studies have found that students using ChatGPT significantly outperformed those who did not in creativity (Lee & Chung, 2024; Urban et al., 2024). However, these studies have not investigated whether the solutions are more flexible, fluent and original in solving MCT. As Kattou et al. (2013) noted, flexibility, fluency and originality are the three most important indicators for evaluating mathematical creativity. Therefore, this study hypothesised as follows: H1: College students using ChatGPT generate solutions that are more flexible (H1a), fluent (H1b), and original (H1c) than those solving MCT without ChatGPT.

ChatGPT provides immediate feedback to help students track learning progress in real time, making them feel more confident in solving tasks (Liu et al., 2025). Furthermore, ChatGPT offers personalised guidance based on students' specific needs, enhancing their confidence in task resolution (Li, 2023). Therefore, this study hypothesised as follows: H2: College students using ChatGPT have higher self-efficacy in task resolution than those who do not.

Based on hybrid human-AI regulation theory (Molenaar 2022a, 2022b), working with ChatGPT requires higher metacognitive skills from students, as they not only monitor and evaluate the information provided by ChatGPT but also refine their own ideas using its suggestions. This process requires strong self-regulation and evaluation skills. However, some students using ChatGPT may fail to perform adequate metacognitive monitoring, leading to lower actual performance. Therefore, they may overestimate their performance in solving tasks. This study hypothesised as follows: H3: College students using ChatGPT overestimate their creativity more than those who do not.

One notable advantage of ChatGPT is its ability to provide instant feedback, thereby enhancing students' interest (Xu et al., 2025). As demonstrated by Turmuzi et al. (2025), ChatGPT improves students' interest in mathematics education. Therefore, this study hypothesised as follows: H4: College students using ChatGPT exhibit greater interest in solving MCT.

By breaking down complex tasks into smaller, more manageable components, ChatGPT provides a clearer execution path. This not only reduces students' time costs but also efficiently handles repetitive tasks, alleviating cognitive load (Simkute et al., 2024). In this way, ChatGPT allows students to focus more on the core issues, reducing task difficulty. Therefore, this study hypothesised as follows: H5: College students using ChatGPT perceive the task as easier.

Finally, based on the guidance of problem cues, ChatGPT can effectively summarise fragmented information, reducing students' mental effort (Steiss et al., 2024). Also, since the assumption is that students using ChatGPT perceive the task as easier, and task difficulty serves as a cue for mental effort allocation (Scheiter et al., 2020). Therefore, we hypothesised as follows: H6: College students using ChatGPT invest less mental effort in solving MCT.

Methods

Procedure

This study was conducted in Yangzhou University, China, before the end of the first semester of 2024. The corresponding author coordinated with counsellors to recruit students from all majors and announced the topic of the study ("Mathematical Creativity Tasks"). Students who volunteered for the study were rewarded with a few credits. The experimental procedure is shown in Figure 1. All students were randomly assigned to either the experimental group (task resolution with ChatGPT) or the control group (task resolution without ChatGPT). Students who withdrew from the study were not included in the data analysis, and their questionnaires were not collected. Therefore, none were excluded from the collected questionnaires. The study took place in a quiet classroom with a maximum of 20 participants at one time (10 participants in each group). The participants worked in separate areas and could not see each other.

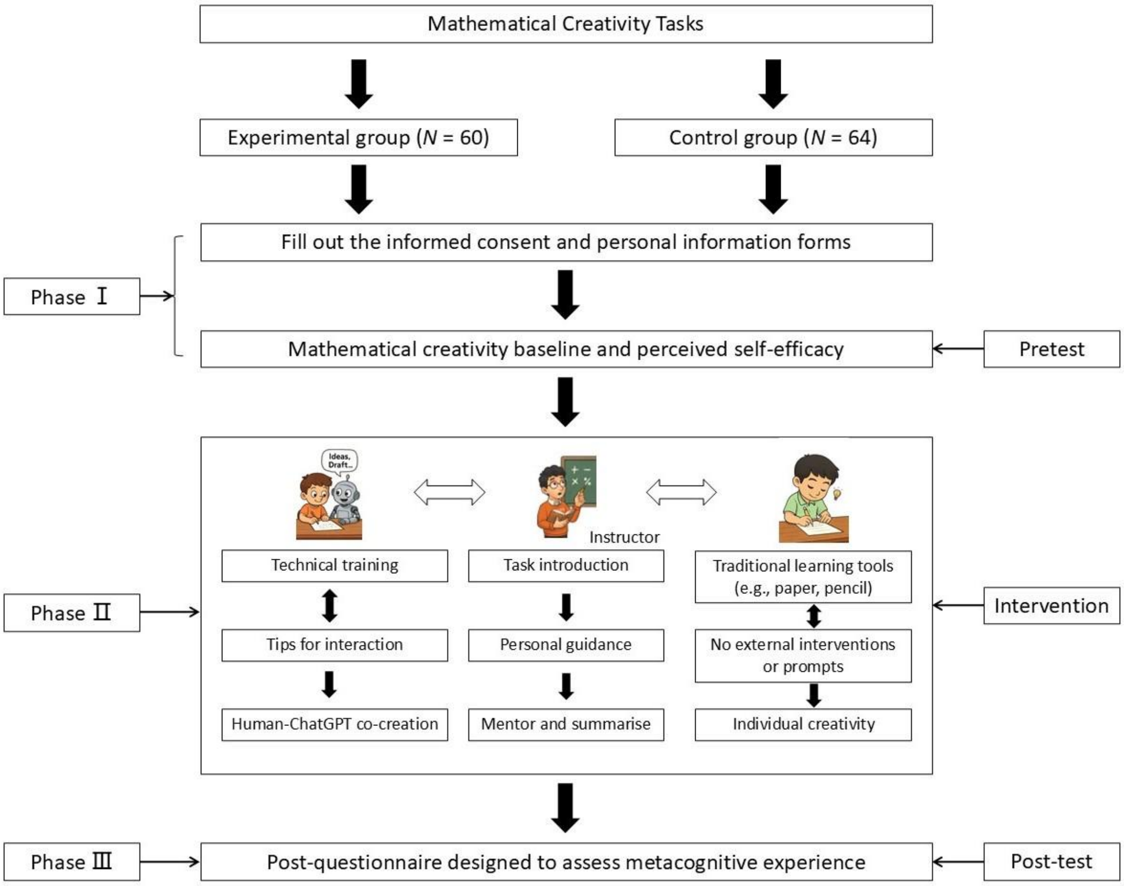


Figure 1. Experimental procedure

In the experimental group, students used personal computers with identical configurations. Although they were free to use a browser for information searches, no one chose to do so. They opened only the required ChatGPT through a single browser tab (Figure 2) and were told that (a) ChatGPT could respond to prompts; (b) providing more background information could generate more accurate answers; (c) users could suggest modifications to refine the answer; and (d) users could upload documents for processing. In the control group, students used only traditional learning tools (e.g., paper and pencil) to solve MCT. They received no support, and they relied on their own creativity in the entire problem-solving process. The experiment lasted 72 minutes (*SD* = 7) in the experimental group and 60 minutes (*SD* = 5) in the control group.

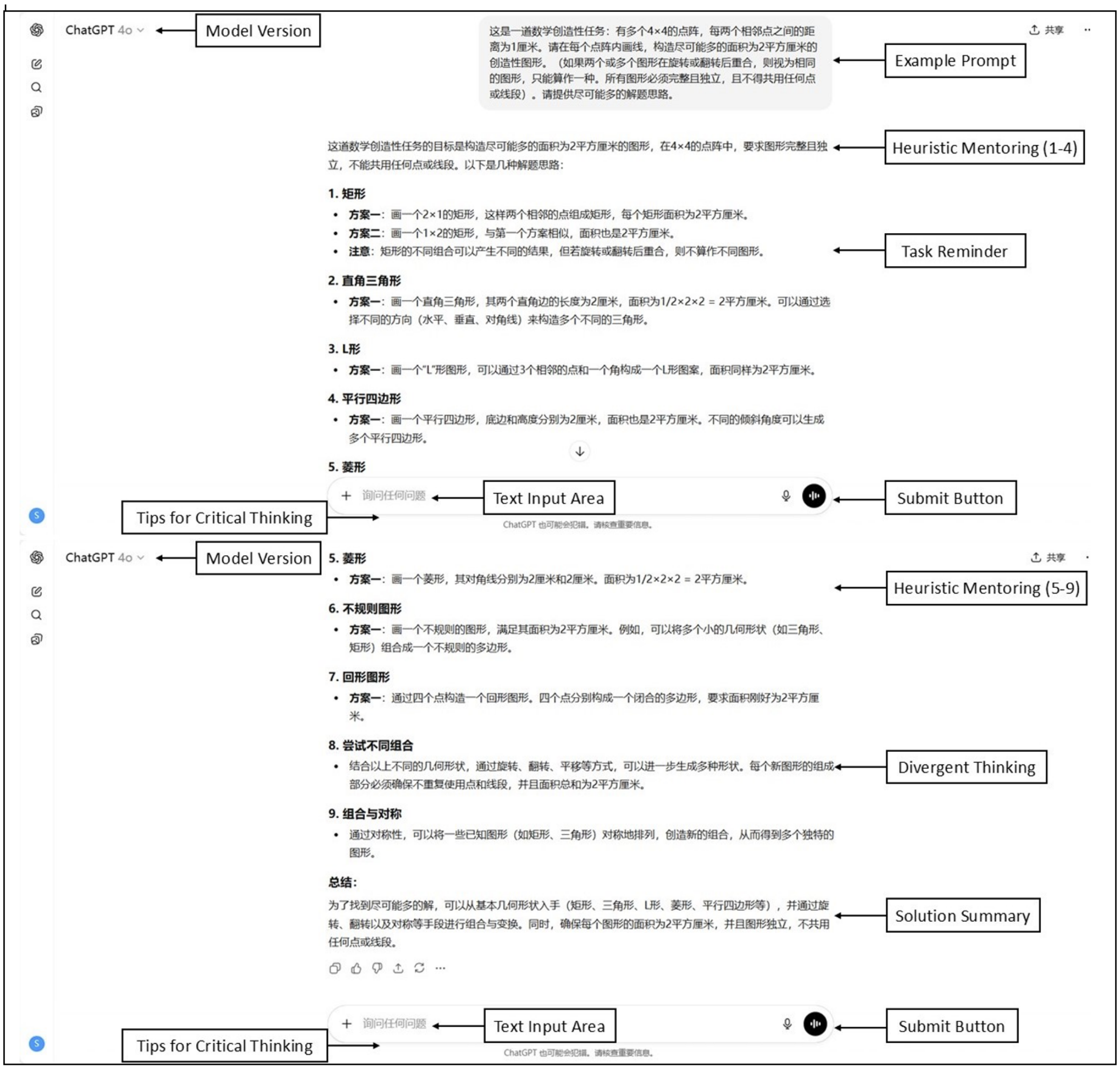


Figure 2. Human-ChatGPT interaction interface

As shown in Table 1, the designed prompts aim to guide students using ChatGPT to solve MCT step by step, stimulating their mathematical creativity. Each prompt was heuristic and gradually increased in complexity. ChatGPT provided relevant feedback based on the students' input, ensuring their continuous progress in solving MCT. All human-ChatGPT interactions were automatically recorded via the platform. The interaction logs included students' input prompts, ChatGPT's responses and the interaction time, and were sequentially numbered for subsequent analysis and verification. Students were allowed to use only the functions provided by ChatGPT and were not permitted to access external resources. Also, students could not change the ChatGPT model or use other chat libraries within the model.

Table 1
Prompt design and its purpose

Prompts	Examples	Purpose
Example prompt	Here is an MCT (e.g., Task 1). Please provide as many solutions as possible.	The goal is not to give direct answers, but to provide additional ideas and encourage students to engage more deeply in problem-solving.
Heuristic mentoring	Try constructing right-angled triangles with different side lengths.	Guide students to broaden their creative thinking by varying the side lengths.
Task reminder	If a figure overlaps after rotation or inversion, it is considered the same and is only counted once.	Ensure that students focus on the task requirements and avoid deviating from the task goal.
Tips for critical thinking	ChatGPT may also make mistakes. Please check important information carefully.	
Divergent thinking	Try combining different geometric shapes.	Encourage students to explore the possibility of generating new shapes by transforming known shapes.

Participants

Since no prior sample size calculation was performed, a sensitivity analysis was conducted based on the sample size (n = 124) and the observed effect sizes ($\eta^2 p$ = 0.04 to 0.16). The results indicated that this study had adequate statistical power (> 0.80) to detect medium to large effects. However, the study may have been underpowered to detect small effects, which should be considered when interpreting null findings.

As shown in Table 2, there were no significant differences in age, gender, major and prior ChatGPT experience – the measurement was derived from Tang et al. (2024) and Urban et al. (2024) – between the two groups. Additionally, all students had no specific mathematics background and had achieved a similar level of performance in prior mathematics courses. Therefore, they were basically homogeneous in mathematical performance. This study was approved by the ethics committee and adhered to the relevant ethical principles.

Table 2
Demographic information

Variable	Experimental	Control	Comparison
Age	M_{age} = 19.20 (SD = 1.36)	M_{age} = 19.20 (SD = 1.10)	$t(122) = 0.01$, $p = 0.989$
Gender			$\chi 2(1) = 2.85$, $p = 0.092$
Male	39	32	
Female	21	32	
Major			$\chi 2(2) = 4.88$, $p = 0.087$
Sciences	16	20	
Liberal arts	16	26	
Engineering	28	18	
Prior ChatGPT experience			$\chi 2(4) = 0.57$, $p = 0.967$
Never	2	2	
Seldom	10	9	
Sometimes	24	28	
Often	23	23	
Always	1	2	

Data analysis

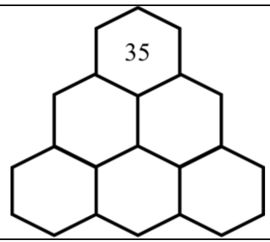
First, the independent-sample *t* tests were conducted to examine the differences between the two groups. To further test H1–6, prior ChatGPT experience was used as a covariate in the analysis of covariance (ANCOVA), aiming to investigate the effect of ChatGPT regardless of prior experience. Before conducting ANCOVA, data sets were screened for outliers, and assumptions including normality, homogeneity of variance, linearity and homogeneity of regression slopes were tested and confirmed to be met. To control multiple comparisons and reduce the risk of false positives, the *p* values for the six ANCOVA were adjusted using the false discovery rate correction (Benjamini-Hochberg method), at a significance level of $\alpha = 0.05$. According to Cohen (1988), effect sizes (η^2p) of 0.01, 0.06 and 0.14 represent small, moderate and large effects, respectively.

Measures

Mathematical creativity baseline

As shown in Table 3, the baseline task adapted from Kattou et al. (2013) was used to measure all students' initial mathematical creativity.

Table 3
Baseline task

In the number pyramid, each regular hexagonal cell must contain a number, and each number can be computed by performing the same operation on the two numbers below it. When filling the pyramid, ensure that number 35 is at the top. Please find as many original solutions as possible.	

MCT

The MCT was adapted from Lee and Seo (2003) to measure mathematical creativity. The initial MCT included five mathematical open-ended tasks. Notably, one task (regular hexagon problem) was excluded from the adapted version due to its low difficulty index (García-García et al., 2024). Salazar Tornel et al. (2022) also argued that it could not differentiate students' levels of mathematical creativity, as it lacked complexity. To better assess mathematical creativity, four more challenging tasks were chosen. The adapted version underwent internal consistency testing, yielding results that demonstrated very high reliability. It includes the following tasks (as shown in Figure 3):

- Task 1: Draw as many 2 cm^2 figures as possible within a 4 × 4 grid (where the distance between adjacent points is 1 cm). If a figure overlaps after rotation or inversion, it is considered the same and only be counted once. All figures should be complete and independent, without sharing any points or line segments.
- Task 2: Devise as many methods as possible to measure the degree of dispersion of five marbles thrown by three students.
- Task 3: When a rectangular prism containing water is tilted, the geometric shape formed at the interface between the water surface and the flask surface also changes. Students should identify as many properties as possible related to the size and shape of the water.
- Task 4: It includes eight geometries, and the goal is to identify as many shapes as possible that share the same features as shape A, and to describe these common characteristics.

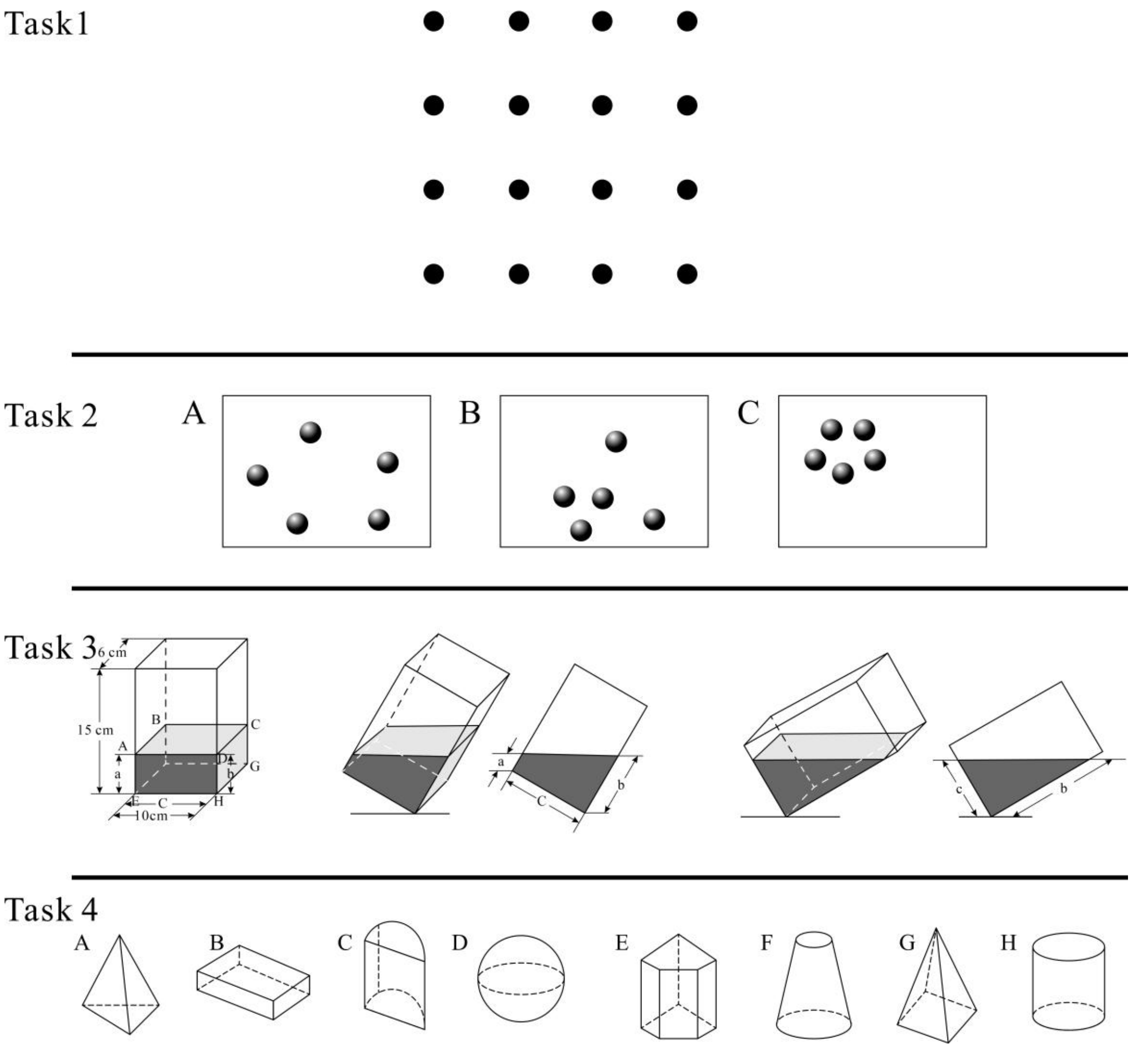


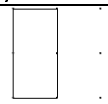
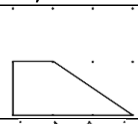
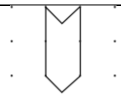
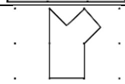
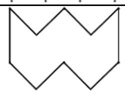
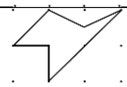
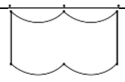
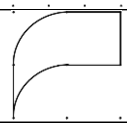
Figure 3. Images of MCT

Mathematical creativity scoring

The assessment of mathematical creativity was based on the flexibility, fluency and originality of students' solutions (Kattou et al., 2013; Lee & Seo, 2003). Flexibility refers to the number of answer categories in each task. Fluency refers to the number of correct answers in each category, with a maximum of 5 points. Originality refers to the rarity of an answer. The proportion of each answer category to the number of total responses was calculated. If the proportion of an answer category fell within 0, 1%; 1%, 2%; or 2%, 3%, the scores were 3, 2 and 1, respectively.

In Task 1, all correct solutions were classified based on Lee and Seo's (2003) scoring table (see Table 4). Then, the originality score for each category was calculated based on its proportion. The individual originality score was calculated by classifying students' solutions and summing the originality scores for each category, with each category counted only once. The individual flexibility score was based on solution categories, while the fluency score was calculated by summing the number of solutions for each category.

Table 4
Task 1 solution classification table

Classification	Examples	
	Symmetrical	Unsymmetrical
Using a single basic figure		
Using more than two figures		
Using the middle point		
Using curved lines		

Therefore, we first marked the solutions in the baseline task and the MCT as correct (1) or incorrect (0). Then, we categorised these correct solutions. Finally, based on the scoring criteria, we assigned scores for the flexibility, fluency and originality of solutions in each task. The scores of the three indicators were summed to measure individual baseline score. For the MCT, the total scores for three indicators were obtained by summing the scores from the four tasks.

Two scorers independently scored the whole data set, with high inter-rater reliability (Cohen's kappa: $k_{task\ 1} = 0.87$, $k_{task\ 2} = 0.83$, $k_{task\ 3} = 0.90$, $k_{task\ 4} = 0.89$). We reanalysed all discrepancies until 100% agreement was reached.

Self-efficacy

After first reading the MCT but before task resolution, students in the control group rated four statements (e.g., "I am confident in measuring the extent of marble scatter" (Puente-Díaz et al. 2021) on a 100-point scale from 1 (*completely disagree*) to 100 (*completely agree*). Students in the experimental group rated eight statements, with four unrelated to ChatGPT (e.g., Even without ChatGPT, I am confident in measuring the extent of marble scatter) and four related to ChatGPT (e.g., "I am confident in using ChatGPT to measure the extent of marble scatter"). The reliability of these statements was strong, Cronbach's $\alpha_{control} = 0.90$, $\alpha_{no\text{-}ChatGPT} = 0.96$, $\alpha_{with\text{-}ChatGPT} = 0.95$.

Self-evaluation

After task resolution, all students evaluated the three indicators of their solutions on a 1–100 scale (Karwowski et al., 2019). Specifically, the statement for flexibility was as follows: "Please rate the variety of solutions you considered in solving MCT". The statement for fluency was as follows: "Please rate the number of correct answers in each solution". The statement for originality was as follows: "Please rate the rarity of your solutions".

Accuracy of self-evaluation (bias index)

The bias index measures the gap between self-evaluation and actual performance (Urban & Urban, 2023a). It was calculated by dividing the difference between the self-evaluation score and the actual performance (converted to a 1–100 scale) by 100. It ranges from -1 to 1, where negative values represent underestimation of creativity, positive values represent overestimation of creativity and values close to zero represent accurate self-evaluation.

Perceived interest

After task resolution, students assessed task interest by answering the following statement (Schiefele, 2009): "Please rate your interest in solving MCT on a scale of 1 (*very boring*) to 100 (*very interesting*)" .

Perceived difficulty

After task resolution, students assessed task difficulty by answering the following statement (Efklides, 2006): "Please rate the difficulty of MCT on a scale from 1 (*very easy*) to 100 (*very difficult*)" .

Perceived mental effort

After task resolution, students assessed their invested mental effort by answering the following statement (Paas, 1992): "Please rate the mental effort you put into solving MCT on a scale from 1 (*no effort*) to 100 (*full effort*)".

Perceived usefulness of ChatGPT

In the experimental group, students rated the usefulness of ChatGPT by answering the final statement (Urban et al., 2024): "Please rate the usefulness of ChatGPT in solving MCT on a scale from 1 (*very useless*) to 100 (*very useful*)".

Results

Table 5 presents the results of the independent-sample *t* tests. The following section presents the results of the ANCOVA.

Table 5
The results of the independent-sample t tests

Variable	Experimental		Control		*t*(122)	*d*	*p*
	M	*SD*	*M*	*SD*			
1. Creativity baseline	15.57	5.77	15.19	6.65	0.34	0.06	0.736
2. Self-efficacy (no ChatGPT)	47.85	20.75	43.78	21.91	1.06	0.19	0.291
3. Self-efficacy (with ChatGPT)	56.93	20.88	-	-	-	-	-
4. MCT (flexibility)	15.12	3.71	12.50	3.49	4.05	0.73	0.000
5. MCT (fluency)	29.25	8.65	23.73	8.08	3.67	0.66	0.000
6. MCT (originality)	4.82	2.64	3.78	2.29	2.34	0.42	0.021
7. Self-evaluation (flexibility)	74.77	13.90	74.64	11.75	0.06	0.01	0.957
8. Self-evaluation (fluency)	80.08	10.83	82.27	12.10	1.06	0.19	0.293
9. Self-evaluation (originality)	73.05	14.51	59.73	18.97	4.37	0.79	0.000
10. Bias index(flexibility)	0.12	0.22	0.15	0.19	0.93	0.15	0.355
11. Bias index (fluency)	0.26	0.20	0.30	0.16	1.10	0.22	0.272
12. Bias index (originality)	0.29	0.29	0.22	0.29	1.42	0.24	0.159
13. Interest	81.17	9.08	73.27	10.82	4.39	0.79	0.000
14. Difficulty	68.65	12.20	75.94	12.50	3.28	0.59	0.001
15. Effort	69.92	12.02	79.14	9.66	4.72	0.85	0.000
16. Usefulness of ChatGPT	65.40	10.44	-	-	-	-	-

Note. The 14 *p* values in the table have been corrected using false discovery rate correction (Benjamini-Hochberg) to control multiple comparisons, with a significance level of $\alpha = 0.05$.

ChatGPT positively influences the flexibility, fluency and originality of solutions (H1a-c)

ChatGPT had a strong effect on flexibility, $F(1,121) = 16.60$, $\eta^2p = 0.12$, $p < 0.001$, and fluency of solutions, $F(1,121) = 13.90$, $\eta^2p = 0.10$, $p < 0.001$, and a moderate effect on originality, $F(1,121) = 5.44$, $\eta^2p = 0.04$, *p*

= 0.021. Prior ChatGPT experience had no impact on flexibility, $F(1,121) = 1.19$, $\eta^2p = 0.01$, $p = 0.278$; fluency, $F(1,121) = 2.66$, $\eta^2p = 0.02$, $p = 0.106$; and originality, $F(1,121) = 0.07$, $\eta^2p = 0.00$, $p = 0.794$.

ChatGPT enhances self-efficacy in solving MCT (H2)

As shown in Figure 4, there was no significant difference in self-efficacy between the two groups when solving MCT without ChatGPT, $F(1,121) = 1.15$, $\eta^2p = 0.01$, $p = 0.285$. However, compared to the control group, students using ChatGPT showed moderately higher self-efficacy in solving MCT, $F(1,121) = 12.15$, $\eta^2p = 0.09$, $p < 0.010$. Interestingly, prior ChatGPT experience was unrelated to self-efficacy for task resolution, $F(1,121) = 3.21$, $\eta^2p = 0.03$, $p = 0.075$.

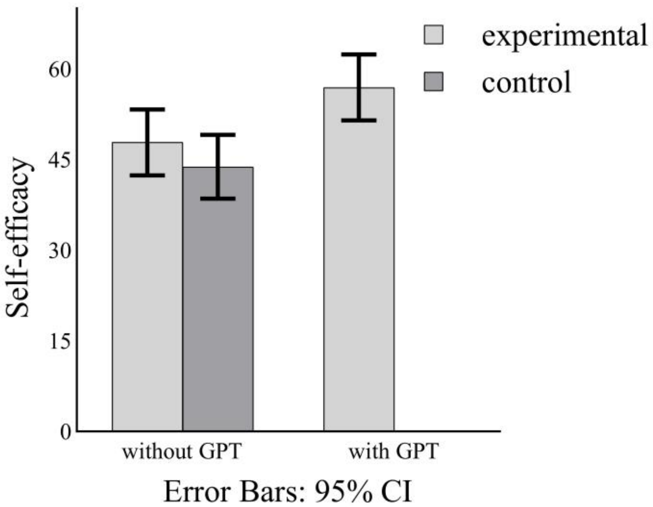


Figure 4. Self-efficacy in solving MCT with and without ChatGPT

Furthermore, students in the experimental group were asked to evaluate their self-efficacy in solving MCT with and without ChatGPT. The moderate within-subject differences suggest that students using ChatGPT exhibited greater confidence in solving MCT than without it, $F(1,117) = 5.70$, $\eta^2p = 0.05$, $p < 0.05$.

From an exploratory perspective, self-efficacy for task resolution and the flexibility, fluency, and originality of solutions had a very small correlation in the experimental group (median: $r_{no\text{-}ChatGPT} = 0.09$, $r_{with\text{-}ChatGPT} = 0.09$) (Table 6). This finding suggests that higher self-efficacy may not be beneficial for mathematical creativity.

ChatGPT makes students overestimate their mathematical creativity (H3)

Interestingly, the accuracy of self-evaluation in flexibility, $F(1,121) = 0.91$, $\eta^2p = 0.01$, $p = 0.342$, fluency, $F(1,121) = 1.26$, $\eta^2p = 0.01$, $p = 0.264$, and originality, $F(1,121) = 2.00$, $\eta^2p = 0.02$, $p = 0.159$, did not show significant differences between the two groups. Also, prior ChatGPT experience had no effect on self-evaluation accuracy of flexibility, $F(1,121) = 1.41$, $\eta^2p = 0.01$, $p = 0.237$; or fluency, $F(1,121) = 0.86$, $\eta^2p = 0.01$, $p = 0.357$; or originality, $F(1,121) = 0.05$, $\eta^2p = 0.00$, $p = 0.818$.

However, the exploratory analysis of correlations among variables offers one important cue. In the experimental group, there was no correlation between students' self-evaluation and actual performance (median $r = -0.06$), showing that students at the group level were unable to calibrate their judgement based on their performance (Table 6). Also, the perceived usefulness of ChatGPT was strongly correlated with higher bias index (median $r = 0.42$). This may be important because the perceived usefulness of ChatGPT was not related to actual performance (median $r = -0.07$). In other words, the more useful the students found ChatGPT, the more they overestimated their performance
.

Table 6

Correlations of variables in the experimental group

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Prior ChatGPT experience																
2. Creativity baseline	0.18															
3. Self-efficacy (no ChatGPT)	-0.02	0.15														
4. Self-efficacy (with ChatGPT)	0.02	0.09	0.08													
5. MCT (flexibility)	-0.02	-0.06	**0.09**	**0.09**												
6. MCT (fluency)	0.17	0.03	0.13	**0.09**	0.40**											
7. MCT (originality)	0.01	-0.13	0.03	-0.07	0.59**	0.12										
8. Self-evaluation (flexibility)	-0.03	-0.04	0.01	0.23	**-0.08**	0.26*	-0.17									
9. Self-evaluation (fluency)	0.05	-0.01	-0.11	0.18	-0.13	**-0.02**	-0.01	0.43**								
10. Self-evaluation (originality)	-0.03	-0.05	-0.21	0.16	0.05	0.13	**-0.06**	0.71**	0.41**							
11. Bias index (flexibility)	-0.01	0.02	-0.06	0.08	-0.77**	-0.12	-0.53**	0.71**	0.37**	0.42**						
12. Bias index (fluency)	-0.11	-0.03	-0.17	0.03	-0.40**	-0.84**	-0.10	0.02	0.57**	0.11	0.30*					
13. Bias index (originality)	-0.02	0.09	-0.13	0.14	-0.47**	-0.03	-0.87**	0.50**	0.21	0.55**	0.66**	0.14				
14. Interest	0.12	0.08	0.01	0.11	0.22	0.13	0.20	0.34**	0.31*	0.27*	0.07	0.07	-0.03			
15. Difficulty	-0.19	0.04	0.08	0.11	0.19	0.12	0.18	0.31*	0.12	0.06	0.07	-0.04	-0.12	0.51**		
16. Effort	-0.11	0.14	0.06	0.14	-0.03	0.19	-0.22	0.47**	0.12	0.41**	0.33*	-0.10	0.39**	0.32*	0.38**	
17. Usefulness of ChatGPT	0.12	0.16	-0.12	0.19	**-0.07**	0.16	-0.19	0.58**	0.34**	0.60**	**0.42****	0.06	0.46**	0.25	0.27*	0.60**

Note. *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

ChatGPT makes MCT more interesting, easier and requires less effort (H4–6)

Students using ChatGPT found MCT more interesting, $F(1,121) = 19.13$, $\eta^2p = 0.14$, $p < 0.001$, and moderately easier, $F(1,121) = 11.05$, $\eta^2p = 0.08$, $p < 0.01$, and they required highly less mental effort, $F(1,121) = 22.18$, $\eta^2p = 0.16$, $p < 0.001$, than those without it (Figure 5).

However, prior ChatGPT experience was unrelated to perceived task interest, $F(1,121) = 0.02$, $\eta^2p = 0.00$, $p = 0.887$; perceived task difficulty, $F(1,121) = 1.98$, $\eta^2p = 0.02$, $p = 0.162$; or the mental effort invested, $F(1,121) = 0.05$, $\eta^2p = 0.00$, $p = 0.830$.

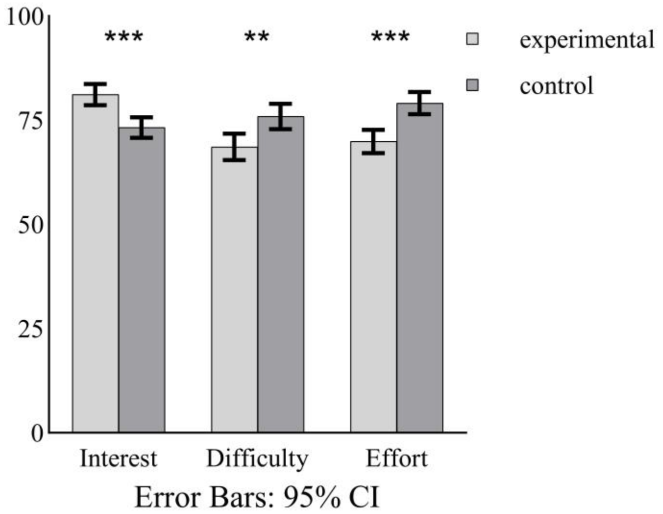


Figure 5. Perceived interest, difficulty and invested mental effort

Discussion

Human–ChatGPT co-creation positively influences mathematical creativity (RQ1, H1a-c)

The study found that college students using ChatGPT generated solutions that exhibited more mathematical creativity than those without it. This finding contradicts Peters et al. (2024), who argued that ChatGPT could not compete with human creativity. One explanation for this discrepancy is the different roles of ChatGPT. In Peters et al.'s work, ChatGPT was treated as a passive creator, with limited human input or refinement. Therefore, its creativity was evaluated in isolation, without human inspiration, which may have limited its performance. In contrast, this study positioned ChatGPT as a co-creator, where students were encouraged to iteratively adjust prompts and actively engage in refining the outputs. This aligns with the hybrid human-AI regulation theory (Molenaar, 2022a, 2022b), which emphasises the importance of self-regulation in co-creation with ChatGPT. Students should not simply rely on ChatGPT to generate ideas; instead, they should actively interact with it and adjust its suggestions. After generating initial prototypes, the process of iteration and reflection is crucial for fostering creativity. This is supported by Urban et al. (2024), who found that the goal of ChatGPT should not be to override students' agency but to help them develop their own ideas. Tang et al. (2024) have also confirmed the potential of human-ChatGPT co-creation in enhancing creativity. However, their finding is not directly comparable to this study, as they focused on pairs and used general creative tasks. Additionally, the experimental group in this study received structured prompts, which makes it difficult to directly compare with previous findings. Ideally, a third condition with human-delivered prompts (without ChatGPT) would have been set to more clearly isolate AI effects.

Based on the current findings, several explanations can be proposed for why ChatGPT positively influenced three indicators of mathematical creativity: flexibility, fluency and originality. ChatGPT offers multiple prototypical solutions for students to choose, helping them iterate quickly and reducing time costs (Hitsuwari et al., 2023). Students can modify and expand upon the reasonable components of these solutions to increase their flexibility. Moreover, ChatGPT is highly effective at generating ideas that are incrementally new rather than radically new (Lee & Chung, 2024). This allows it to proficiently combine various elements from existing solutions. For instance, it can deconstruct composite figures in Task 1 into basic elements and then recombine them randomly, thereby enhancing the fluency of solutions. Finally, ChatGPT can integrate interdisciplinary knowledge, encouraging students to explore unconventional yet mathematically sound solutions. As suggested by Song et al. (2025), human-ChatGPT co-creation may inspire novel combinations of concepts, generating more original final solutions.

Prior ChatGPT experience has no effect on human–ChatGPT co-creation

Zamfirescu-Pereira et al. (2023) claimed that experienced users often have smoother human–AI interactions, as inexperienced users struggle with unclear prompts when interacting with them, resulting in task failure or difficulties in assessing prompt quality. However, in solving MCT, the convenience and efficiency provided by prior use of ChatGPT may lead experienced students to depend heavily on it, thereby diminishing their metacognitive regulation skills. In contrast, inexperienced students interact with ChatGPT more cautiously. They will more accurately monitor the interaction process and continuously adjust the co-created outcomes. The novelty effect also encourages them to use it as an exploratory tool, relying more on feeding their own prototypical solutions to it (Liu et al., 2025). Furthermore, the current version of ChatGPT has addressed the shortcomings of unclear prompts (e.g., uploading images or attachments), enabling all users to gain effective inspiration regardless of their prior experience.

While ChatGPT enhances self-efficacy, higher self-efficacy may not benefit creativity (RQ2, H2)

When aggregated for all students, higher self-efficacy is associated with better mathematical creativity. This can be explained by Schunk and Mullen (2012), who reported that students with higher self-efficacy tend to focus their attention on achieving their goals and seeking solutions from various perspectives. Conversely, those with lower self-efficacy lack the confidence to develop creative solutions, assuming that the answers are already known. However, at the within-subject level, higher self-efficacy is negatively correlated with performance (Bicer et al., 2020), as students believe they will perform better on certain tasks than others. This belief may hinder their performance on tasks they believe they will excel at. Since ChatGPT is an emerging technology that enhances self-efficacy at the between- and within-subject levels, it may be that these improvements in self-efficacy have yet to be fully considered in solving the MCT and self-evaluating the outcomes. In other words, the increase in self-efficacy from ChatGPT may be making students overconfident in their abilities, which could negatively impact their performance.

ChatGPT may hurt the accuracy of self-evaluation (RQ3, H3)

This study revealed a paradox: although ChatGPT significantly enhanced college students' mathematical creativity, they tended to overestimate their own performance. According to the hybrid human-AI regulation theory (Molenaar, 2022a, 2022b), effective human–AI collaboration depends not only on the quality of AI-generated outputs but also on students' ability to monitor, evaluate and regulate their cognition throughout human–ChatGPT co-creation process. However, ChatGPT may inadvertently reduce students' engagement in such self-regulatory processes. High-quality output may increase reliance on ChatGPT, reducing students' metacognitive monitoring and weakening their accuracy of self-evaluation. Therefore, students' reliance on AI-generated output without sufficient metacognitive regulation leads to inaccurate self-evaluation.

Although the study indicated that ChatGPT is positively associated with mathematical creativity, it remains worth exploring whether these improvements reflect genuine creative development or merely performance gains resulting from reliance on ChatGPT. According to Molenaar's (2022a, 2022b) hybrid human-AI regulation theory, the development of creativity depends on students' ability to monitor, evaluate and regulate the human–ChatGPT co-creation process, rather than outsourcing cognitive work to ChatGPT. Therefore, genuine creative development means that, when co-creating with AI, students are not only able to actively monitor, evaluate and regulate AI-generated ideas but also gradually internalise these metacognitive skills and transfer them to new situations, thereby developing stable creative mathematical problem-solving ability. In contrast, if students neglect metacognitive skills, improvements in mathematical creativity are more likely to reflect performance gains resulting from reliance on ChatGPT. Such gains are limited to immediate tasks and are difficult to transform into sustainable competence. As identified by Zamfirescu-Pereira et al. (2023), inexperienced users struggled with self-monitoring during AI-assisted problem solving, leading to cognitive passivity. Future work could use lag sequential analysis to code the behaviour sequences of students in the experimental group (Ren et al., 2026; Zhou et al., 2025). For example, frequent explanation and evaluation suggest that students actively adjusted AI-generated ideas, which may indicate genuine creative development. In contrast, behavioural patterns characterised by rapid selection and direct adoption of AI-generated ideas may show that students relied more on ChatGPT for performance gains. Future longitudinal studies are needed to assess whether long-term use of ChatGPT fosters the internalisation of creative thinking or leads to dependency. Research should also examine whether students take responsibility for monitoring, evaluating and regulating AI-generated ideas, which is regarded as a sign of genuine creative development.

ChatGPT affects metacognition in human–ChatGPT co-creation (RQ3, H4-6)

The explorations provide three heuristic cues about students using ChatGPT (see Table 6). First, there was a moderate correlation between the perceived usefulness of ChatGPT and overestimation of mathematical creativity. Second, the perceived task interest was positively correlated with self-evaluation. Third, the perceived task difficulty and interest were positively related to invested mental effort. Based on these findings, it can be inferred that when students perceive ChatGPT as more useful, they are likely to overestimate their performance. This overestimation may lead to greater satisfaction with their current solutions, thereby increasing their interest in solving tasks. Meanwhile, if students find the task more interesting, they may invest greater effort, even if they perceive the task as somewhat challenging. However, with cross-sectional correlations, directionality cannot be determined. Another explanation is that students who overestimated their performance rated ChatGPT as more useful (reverse causality) and the MCT as more interesting, leading them to invest more effort, even if they perceive the MCT as challenging. Notably, there was no correlation between students' self-evaluation and their actual creativity. In other words, students believe they are performing well and exerting greater effort, yet this may not translate into actual performance, as effectiveness can be influenced by cognitive biases or other factors (Scheiter et al., 2020). Therefore, future work should explore how to enhance mathematical creativity while avoiding overestimation. It is essential to explore how to transform students' interests and efforts into actual mathematical creativity.

Limitations and future work

This study has several limitations. First, the study employed MCT that were influenced by some factors (e.g., mathematical beliefs and motivations), which could potentially impact their mathematical creativity. Although known differences between the groups were controlled through pretests and covariates, these methods did not account for all potential confounding variables. Therefore, while the intervention effect of ChatGPT was observed, these unmeasured psychological factors may still serve as potential drivers of the observed differences. Future research should strengthen experimental designs by incorporating more comprehensive baseline measurements, using advanced matching techniques to create more comparable groups, and employing multiple control groups to identify potential confounders.

Second, using a 5-point scale to measure prior ChatGPT experience may not fully capture the complexity of their ChatGPT experience. This measurement did not account for the depth or specific context of use, oversimplifying students' familiarity and expertise with ChatGPT. Future work could incorporate more detailed behavioural measurements or context-specific assessments of prior experience to better reflect the applications of ChatGPT in real learning environments.

Moreover, the MCT used in this study focused on geometry, which limits the generalisability of the findings to other mathematical areas. Also, this study was conducted under controlled ideal experimental conditions, which may not fully reflect real classroom practices. Future research should develop comprehensive MCT covering a broader range of mathematical areas and conduct research under real classroom conditions to enhance the ecological validity of the findings.

It is worth noting that the students using ChatGPT took longer to solve the MCT despite their reporting lower difficulty and less mental effort. Students with more time have more opportunities to generate, refine and evaluate solutions. Therefore, future studies should further explore the temporal aspects of a problem-solving process by analysing trace data and should consider time as a covariate in sensitivity analyses.

Finally, this study was conducted at a single university in China, which may limit the generalisability of the findings to other cultural contexts and educational levels. Future work should include participants from different cultural backgrounds and expand the educational levels to enhance the external validity of the results. Furthermore, the students who volunteered to earn credit may introduce self-selection bias, as they may differ from non-participating students in terms of motivation or interest in AI. Future research should consider using random sampling methods to reduce this bias.

Theoretical and practical implications

From a theoretical perspective, this study validates Molenaar's (2022a, 2022b) hybrid human-AI regulation theory by showing that metacognitive regulation is crucial for human-AI collaboration in solving MCT. In other words, ChatGPT positively affects mathematical creativity while also requiring greater use of metacognitive skills. This study further refined the theory by finding that ChatGPT enhances college students' mathematical creativity but reduces the accuracy of their metacognitive evaluation. Therefore, this study not only applied the hybrid human-AI regulation theory to the ChatGPT context but also refined it in a ChatGPT context: improved performance and reduced metacognitive evaluation may coexist in human-AI collaboration. This study also contributes to the growing body of literature on the impact of ChatGPT on mathematical creativity, providing a foundation for discussing its application in higher education. Additionally, college students should consciously use effective monitoring cues in solving MCT and self-evaluating, providing a new analytical framework for understanding the impact of ChatGPT on mathematical creativity. However, these findings may not generalise to younger students, non-geometry domains or students who are required (rather than choose) to use ChatGPT.

From a practical perspective, policymakers should incorporate human–ChatGPT co-creation into national curriculum. Specifically, AI literacy modules can be embedded within existing courses to teach students to craft effective prompts, critically assess AI-generated outputs and reflect on the co-creation process. Evaluation frameworks should go beyond final outcomes and include students' usage trajectories (e.g., submission of pre- and post-AI versions of their work, prompting logs) to assess whether they are engaging in genuine creative iteration or simply accepting AI suggestions.

Developers should design ChatGPT platforms that support self-monitoring and reflection. This could involve visual dashboards (track the number and quality of prompt revisions), embedded reflective prompts (e.g., why did you choose this solution?) or feedback mechanisms (compare their own ideas with AI-generated ones). We recommend designing a tiered interface to ensure that students use ChatGPT strategically rather than passively. For example, a beginner mode provides step-by-step guidance for novice users, while an advanced mode allows experienced users to optimise prompts more effectively.

Educators should adopt a "humans first, AI second" workflow, requiring students to first independently develop an initial solution before using ChatGPT to refine it. Then, students could evaluate the two versions through self-reflection, articulating the improvements and the rationale behind them. Teachers could also introduce structured reflection prompts or evaluation rubrics, requiring students to explain, revise and justify their reasoning. For instance, students could annotate final solutions to indicate which parts generated from ChatGPT and which were their own contributions. Also, classroom activities like peer review or human–ChatGPT collaboration debates could further assess ChatGPT outputs, preventing blind acceptance.

Conclusions

This study found that ChatGPT positively affects college students' mathematical creativity but may influence their ability to evaluate their own work accurately. Therefore, students using ChatGPT should consciously apply metacognitive skills during task resolution and self-evaluation. It is crucial for the future development of hybrid human-AI regulation theory and is important for students using ChatGPT in solving MCT. In pedagogical practice, both humans and ChatGPT should share agency, with each contributing unique strengths to achieve the best results together.

Author contributions

Zhiwei Liu: Conceptualisation, Data curation, Formal analysis, Writing – original draft, Writing – review and editing; **Gang Li:** Data curation, Investigation, Writing – review and editing; **Haode Zuo:** Conceptualisation, Investigation, Formal analysis; **Yongjing Lu:** Writing – review and editing; **Jue Feng:** Writing – review and editing.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 12371140). This study was also supported by the Doctoral Student Program of the Young S&T Talents Cultivation Project, CAST; the Jiangsu Provincial Higher Education Teaching Reform Project, "Construction and Practice of a General Education Curriculum System on Mathematical Principles of Artificial Intelligence for Science and Engineering Undergraduates" (Project Approval No. 2025JGYB800); the 2025 Yangzhou University Humanities and Social Sciences Research Fund Special Project "Research on Teacher Education Talent Cultivation Mechanisms and Feasible Pathways for 'Optimised Combination and Transformational Integration' Facilitated by Artificial Intelligence" (Project Approval No. XJJZX2025-1); the Yangzhou University Graduate Education Teaching Reform Project (Project Approval No. XJGKT25_003); the Sanlian Shuyuan Master Teacher Studio; and the Excellent Teaching Team of the Yangzhou University "Qinglan Project".

References

Bahar, A. K., Can, I., & Maker, C. J. (2024). What does it take to be original? An exploration of mathematical problem solving. *Thinking Skills and Creativity*, *53*, Article 101592. https://doi.org/10.1016/j.tsc.2024.101592

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, *5*, Article 100161. https://doi.org/10.1016/j.caeai.2023.100161

Bicer, A., Lee, Y., Perihan, C., Capraro, M. M., & Capraro, R. M. (2020). Considering mathematical creative self-efficacy with problem posing as a measure of mathematical creativity. *Educational Studies in Mathematics*, *105*(3), 457–485. https://doi.org/10.1007/s10649-020-09995-8

Boussioux, L., Lane, J. N., Zhang, M., Jacimovic, V., & Lakhani, K. R. (2024). *Generative AI and creative problem solving* (Working Paper No. 24-005). Harvard Business School. https://doi.org/10.2139/ssrn.4533642

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review*, *1*(1), 3–14. https://doi.org/10.1016/j.edurev.2005.11.001

Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., Kosch, T., Shen, C., & Welsch, R. (2026). AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior*, *175*, Article 108779. https://doi.org/10.1016/j.chb.2025.108779

García-García, J., Martínez-Artero, M. R. N., & Olivares-Carrillo, P. (2024). Exploring the influence of musical training on mathematical creativity. *Thinking Skills and Creativity*, *52*, Article 101498. https://doi.org/10.1016/j.tsc.2024.101498

Greene, J. A., Freed, R., & Sawyer, R. K. (2019). Fostering creative performance in art and design education via self-regulated learning. *Instructional Science*, *47*(2), 127–149. https://doi.org/10.1007/s11251-018-9479-8

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Habib, S., Vogel, T., Anli, X., & Thorne, E. (2024). How does generative artificial intelligence impact student creativity? *Journal of Creativity*, *34*(1), Article 100072. https://doi.org/10.1016/j.yjoc.2023.100072

Hawlitschek, A., & Joeckel, S. (2017). Increasing the effectiveness of digital educational games: The effects of a learning instruction on students' learning, motivation and cognitive load. *Computers in Human Behavior*, *72*, 79–86. https://doi.org/10.1016/j.chb.2017.01.040

Hitsuwari, J., Ueda, Y., Yun, W., & Nomura, M. (2023). Does human-AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior*, *139*, Article 107502. https://doi.org/10.1016/j.chb.2022.107502

Karwowski, M., Lebuda, I., & Beghetto, R. A. (2019). Creative Self-Beliefs. In I. J. C. Kaufman & R. J. Sternberg (Eds.), The Cambridge handbook of creativity (pp. 396-418). Cambridge University Press.

Kattou, M., Kontoyianni, K., Pitta-Pantazi, D., & Christou, C. (2013). Connecting mathematical creativity to mathematical ability. *ZDM–Mathematics Education*, *45*(2), 167–181. https://doi.org/10.1007/s11858-012-0467-1

Lee, B. C., & Chung, J. (2024). An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour*, *8*(10), 1906–1914. https://doi.org/10.1038/s41562-024-01953-1

Lee, K.-S., & Seo, J.-J. (2003). A development of the test for mathematical creative problem solving ability. *Journal of the Korea Society of Mathematical Education Series D: Research in Mathematical Education, 7*(3), 163–189.

Leikin, R., & Elgrably, H. (2022). Strategy creativity and outcome creativity when solving open tasks: Focusing on problem posing through investigations. *ZDM – Mathematics Education*, *54*, 35–49. https://doi.org/10.1007/s11858-021-01319-1

Leikin, R., & Sriraman, B. (2022). Empirical research on creativity in mathematics (education): From the wastelands of psychology to the current state of the art. *ZDM – Mathematics Education*, *54*, 1–17. https://doi.org/10.1007/s11858-022-01340-y

Levav-Waynberg, A., & Leikin, R. (2012). The role of multiple solution tasks in developing knowledge and creativity in geometry. *The Journal of Mathematical Behavior*, *31*(1), 73–90. https://doi.org/10.1016/j.jmathb.2011.11.001

Li, H. (2023). Effects of a ChatGPT-based flipped learning guiding approach on learners' courseware project performances and perceptions. *Australasian Journal of Educational Technology*, *39*(5), 40–58. https://doi.org/10.14742/ajet.8923

Liu, Z., Zuo, H., & Lu, Y. (2025). The impact of ChatGPT on students' academic achievement: A meta-analysis. *Journal of Computer Assisted Learning*, *41*(4), Article e70096. https://doi.org/10.1111/jcal.70096

Molad, O., Levenson, E. S., & Levy, S. (2020). Individual and group mathematical creativity among post–high school students. *Educational Studies in Mathematics*, *104*(2), 201–220. https://doi.org/10.1007/s10649-020-09952-5

Molenaar, I. (2022a). The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. *Computers and Education: Artificial Intelligence*, *3*, Article 100070. https://doi.org/10.1016/j.caeai.2022.100070

Molenaar, I. (2022b). Towards hybrid human-AI learning technologies. *European Journal of Education*, *57*(4), 632–645. https://doi.org/10.1111/ejed.12527

Niloy, A. C., Akter, S., Sultana, N., Sultana, J., & Rahman, S. I. U. (2024). Is ChatGPT a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning*, *40*(2), 919–930. https://doi.org/10.1111/jcal.12929

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

OpenAI. (2023). *GPT-4 technical report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Peters, M. A., Jackson, L., Papastephanou, M., Jandrić, P., Lazaroiu, G., Evers, C. W., Cope, B., Kalantzis, M., Araya, D., & Tesar, M. (2024). AI and the future of humanity: ChatGPT-4, philosophy and education–Critical responses. *Educational Philosophy and Theory*, *56*(9), 828–862. https://doi.org/10.1080/00131857.2023.2213437

Puente-Díaz, R., Cavazos-Arroyo, J., & Puerta-Sierra, L. (2021). Idea generation, selection, and evaluation: A metacognitive approach. *The Journal of Creative Behavior*, *55*(4), 1015–1027. https://doi.org/10.1002/jocb.505

Ren, W., Li, J., Pi, Z., Guo, J., & Li, X. (2026). How do high-performers and low-performers differently engage in collaborative creative problem solving with a conversational GenAI chatbot? *Thinking Skills and Creativity*, *60*, Article 102134. https://doi.org/10.1016/j.tsc.2026.102134

Ron-Ezra, M., & Levenson, E. S. (2025). Exploring how students with learning disorders engage with open arithmetic tasks using a mathematical creativity lens. *Educational Studies in Mathematics*, *120*(3), 587–608. https://doi.org/10.1007/s10649-025-10426-9

Salazar Tornel, M., Bermejo García, R., & Ferrando, M. (2022). Análisis de una prueba para medir la creatividad matemática [Analysis of a test for measuring mathematical creativity]. *Revista de Estudios e Investigación en Psicologíay Educación*, *9*(1), 76–96. https://doi.org/10.17979/reipe.2022.9.1.8812

Scheiter, K., Ackerman, R., & Hoogerheide, V. (2020). Looking at mental effort appraisals through a metacognitive lens: Are they biased? *Educational Psychology Review*, *32*(4), 1003–1027. https://doi.org/10.1007/s10648-020-09555-9

Schiefele, U. (2009). Situational and individual interest. In K. R. Wentzel (Ed.), *Handbook of motivation at school* (1st ed., pp. 197–223). Routledge.

Schunk, D. H., & Mullen, C. A. (2012). Self-efficacy as an engaged learner. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 219–235). Springer. https://doi.org/10.1007/978-1-4614-2018-7_10

Simkute, A., Tankelevitch, L., Kewenig, V., Scott, A. E., Sellen, A., & Rintel, S. (2024). Ironies of generative AI: Understanding and mitigating productivity loss in human-AI interaction. *International Journal of Human–Computer Interaction*, *41*(5), 2898–2919. https://doi.org/10.1080/10447318.2024.2405782

Song, Y., Kim, J., Xing, W., Liu, Z., Li, C., & Oh, H. (2025). Elementary school students' and teachers' perceptions toward creative mathematical writing with generative AI. *Journal of Research on Technology in Education*, *58*, 1–23. https://doi.org/10.1080/15391523.2025.2455057

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Suherman, S., & Vidákovich, T. (2022). Assessment of mathematical creative thinking: A systematic review. *Thinking Skills and Creativity*, *44*, Article 101019. https://doi.org/10.1016/j.tsc.2022.101019

Suherman, S., & Vidákovich, T. (2024). Role of creative self-efficacy and perceived creativity as predictors of mathematical creative thinking: Mediating role of computational thinking. *Thinking Skills and Creativity*, *53*, Article 101591. https://doi.org/10.1016/j.tsc.2024.101591

Tang, M., Hofreiter, S., Werner, C. H., Zielińska, A., & Karwowski, M. (2024). Who is the best creative thinking partner? An experimental investigation of human-human, human-internet, and human-AI co-creation. *The Journal of Creative Behavior*, *59*(3), Article e1519. https://doi.org/10.1002/jocb.1519

Turmuzi, M., Azmi, S., & Kertiyani, N. M. I. (2025). ChatGPT in school mathematics education: A systematic review of opportunities, challenges, and pedagogical implications. *Teaching and Teacher Education*, *170*, Article 105286. https://doi.org/10.1016/j.tate.2025.105286

Urban, K., & Urban, M. (2023a). How can we measure metacognition in creative problem-solving? Standardization of the MCPS scale. *Thinking Skills and Creativity*, *49*, Article 101345. https://doi.org/10.1016/j.tsc.2023.101345

Urban, M., & Urban, K. (2023b). Do we need metacognition for creativity? A necessary condition analysis of creative metacognition. *Psychology of Aesthetics, Creativity, and the Arts, 19*(6), 1467–1483. https://doi.org/10.1037/aca0000647

Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers & Education*, *215*, Article 105031. https://doi.org/10.1016/j.compedu.2024.105031

Winne, P. H. (2017). Cognition and metacognition within self-regulated learning. In P. A. Alexander, D. H. Schunk, & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 36–48). Routledge. https://doi.org/10.4324/9781315697048-3

Xu, X., Qiao, L., Cheng, N., Liu, H., & Zhao, W. (2025). Enhancing self-regulated learning and learning experience in generative AI environments: The critical role of metacognitive support. *British Journal of Educational Technology*, *56*(5), 1842–1863. https://doi.org/10.1111/bjet.13599

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, & M. L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–21). Association for Computing Machinery. https://doi.org/10.1145/3544548.3581388

Zhou, E., & Lee, D. (2024). Generative artificial intelligence, human creativity, and art. *PNAS Nexus*, *3*(3), Article pgae052. https://doi.org/10.1093/pnasnexus/pgae052

Zhou, R., He, X., Fan, Q., Li, Y., Li, Y., Xiao, X., & Fang, J. (2025). Exploring ChatGPT-facilitated scaffolding in undergraduates' mathematical problem solving. *Journal of Computer Assisted Learning*, *41*(4), Article e70077. https://doi.org/10.1111/jcal.70077

Corresponding author: Haode Zuo, yzzxzhd@yzu.edu.cn

Please cite as: Liu, Z., Li, G., Zuo, H., Lu, Y., & Feng, J. (2026). Exploring the impact of ChatGPT on college students' mathematical creativity: A quasi-experimental study. *Australasian Journal of Educational Technology*, *42*(2), 39–59. https://doi.org/10.14742/ajet.10564