Journal of Interdisciplinary Studies in Education

Volume 15, Issue 4 (2026), pp. 193-220
Journal of Interdisciplinary Studies in Education
ISSN: 2166-2681Print 2690-0408 Online | https://ojed.org/jise
https://doi.org/10.32674/sr0hxm38

Collaborative Learning Interventions for Elementary Students' Mathematical Creative Reasoning: A Meta-Analysis

Tety Nur Cholifah, Budi Eko Soetjipto, Punaji Setyosari, Sri Untari
State University of Malang (Universitas Negeri Malang), Indonesia

ABSTRACT

This meta-analysis synthesizes experimental and quasi-experimental evaluations of collaborative learning interventions for elementary students' mathematical creative reasoning (CMR). After a verification audit, 30 studies (N = 3,108) from 11 countries (1997–2024) met the inclusion criteria. A random-effects model yielded a large effect (Hedges' g = 1.58, 95% CI [1.46, 1.70], p < .001) with moderate heterogeneity (I^2 = 42.1%), interpreted cautiously given that 50% of studies were Indonesian; the non-Indonesian subsample yielded g = 1.49. Open-ended/multiple solution tasks and problem-based learning produced the largest effects; instructional time was not a significant moderator. The findings imply that higher-education teacher preparation should embed evidence-based collaborative pedagogies to ensure curriculum continuity in tertiary mathematics education.

Keywords: collaborative learning, elementary education, mathematical creative reasoning, meta-analysis, teacher education

INTRODUCTION

Despite three decades of research on collaborative learning in elementary mathematics, no quantitative synthesis has specifically targeted *mathematical creative reasoning* (CMR) as the primary outcome, and existing reviews have not interrogated how findings from elementary contexts inform teacher preparation in higher education. The present meta-analysis addresses this gap through three contributions. **First,** it provides the first synthesis of 30 verified experimental and quasi-experimental primary studies isolating CMR rather than general mathematics achievement. **Second,** it offers an explicitly cautious effect-size interpretation, with sensitivity analyses isolating geographic robustness (Indonesian vs. non-Indonesian subsamples) and leave-one-out checks, while transparently disclosing methodological limitations of the available primary study data. **Third,** it translates findings into implications for higher education by analyzing how elementary-level evidence should shape pre-service teacher curricula and tertiary mathematics pedagogy. These contributions distinguish this work from prior elementary-focused syntheses and align it with the journal's scope on higher education research and practice.

In an increasingly competitive job market, where the demand for mathematics-intensive science and engineering jobs is outpacing overall job growth three to one (National Mathematics Advisory Panel, 2008), the ability to think mathematically is a crucial skill. Although the percentage of students reaching proficient levels in mathematics has increased over the past decade, mathematical gains made in elementary school are, on average, not matched in later years (National Mathematics Advisory Panel, 2008). For example, the 2022 Programme for International Student Assessment (PISA) data indicated that Indonesia ranked 68th out of 81 countries in mathematics literacy with a score of 366, far below the OECD average of 472, and only approximately 1% of Indonesian students reached proficiency levels in mathematics involving complex reasoning and creative problem solving (OECD, 2023).

Students' *creative mathematical reasoning,* the ability to create novel problem-solving strategies supported by plausible arguments and grounded in intrinsic mathematical properties (Lithner, 2008), is crucial for mathematical proficiency but is often underdeveloped in traditional classroom settings. Lithner's (2017) framework distinguishes CMR from imitative reasoning through three criteria: novelty, plausibility, and mathematical anchoring. This construct has been validated with elementary students and provides the most widely used approach for assessing mathematical creativity in contemporary research (Leikin & Sriraman, 2022; Schindler & Lilienthal, 2020), consistent with interdisciplinary perspectives that position creative and critical thinking as central, cultivable educational aims (de Gastyne, 2020). Recent work continues to refine the construct in tertiary contexts, demonstrating that the same theoretical foundations apply to

undergraduate mathematics learning, pre-service teacher education, and creative problem solving across diverse domains (Hadar & Tirosh, 2019; Lynch et al., 2025; Maker, 2026; Norqvist et al., 2025).

Collaborative learning has long been recognized as an effective pedagogical approach for developing cognitive and social competencies in students (Slavin, 2015). Social interdependence theory identifies five essential elements of cooperative learning: positive interdependence, individual accountability, promotive interaction, social skills, and group processing (Buchs et al., 2017). Slavin's (2015) Student Team Learning models, STAD, TAI, and Jigsaw, have demonstrated consistent effectiveness across mathematics education contexts (Capar & Tarim, 2015; Talkhan et al., 2025; Turgut & Turgut, 2018).

Several meta-analyses already exist regarding the effectiveness of collaborative learning for students' mathematics achievement; however, these reviews targeted general mathematics achievement (Capar & Tarim, 2015; Kyndt et al., 2013; Turgut & Turgut, 2018) rather than CMR specifically. Capar and Tarim (2015) reported $d = 0.59$ for cooperative learning. Turgut and Turgut (2018) reported $g = 0.84$ in Turkey. Kyndt et al. (2013) reported $d = 0.54$. None of the studies examined CMR as a specific outcome or focused exclusively on elementary students. More recent syntheses have continued to confirm the broad effectiveness of technology-enhanced collaborative learning approaches across educational contexts (Tlili et al., 2025), and recent second-order syntheses confirm consistent moderate-to-large cooperative learning effects on achievement and higher-order thinking (Erdem, 2026); however, they also do not directly address the CMR construct in elementary contexts that feed into tertiary teacher preparation pipelines.

A second limitation of prior reviews is the inconsistency in operationalizing the CMR construct. Researchers use "mathematical creativity," "mathematical creative thinking," and "creative mathematical reasoning" interchangeably despite different theoretical nuances (Leikin & Sriraman, 2022; Nadjafikhah et al., 2012; Sriraman, 2005). Leikin (2009) operationalized mathematical creativity through multiple solution tasks, measuring fluency, flexibility, and originality. A further limitation is that the existing evidence base is heavily clustered in a small number of national contexts, particularly Indonesia, which raises legitimate questions about generalizability that have not been formally addressed in earlier syntheses (Siagian et al., 2023; Zhan et al., 2024).

The most relevant prior meta-analysis (Siagian et al., 2023; Zhan et al., 2024) included only 23 studies and focused on the Indonesian context, while earlier syntheses (Capar & Tarim, 2015) covered 1988–2010, leaving recent research (2015–2025) inadequately integrated. Moderator analyses in previous reviews have not deeply explored collaborative learning type, intervention duration, and measurement-instrument characteristics. Crucially, prior reviews have also failed to translate elementary-level findings into actionable guidance for higher

education, even though pre-service teacher education programs are the primary mechanism through which evidence-based pedagogies reach future elementary classrooms (Cochran-Smith et al., 2018).

To address these gaps, the present review answered five research questions:

RQ1: What is the average effect of collaborative learning interventions on elementary students' mathematical creative reasoning?

RQ2: What is the degree of heterogeneity in effect sizes across studies, and how robust is the pooled estimate?

RQ3: Which instructional and methodological characteristics moderate intervention effectiveness?

RQ4: Is there evidence of publication bias, and how robust are findings to potential bias?

RQ5: How should the synthesized evidence inform pre-service teacher preparation and curriculum continuity in higher education?

METHOD

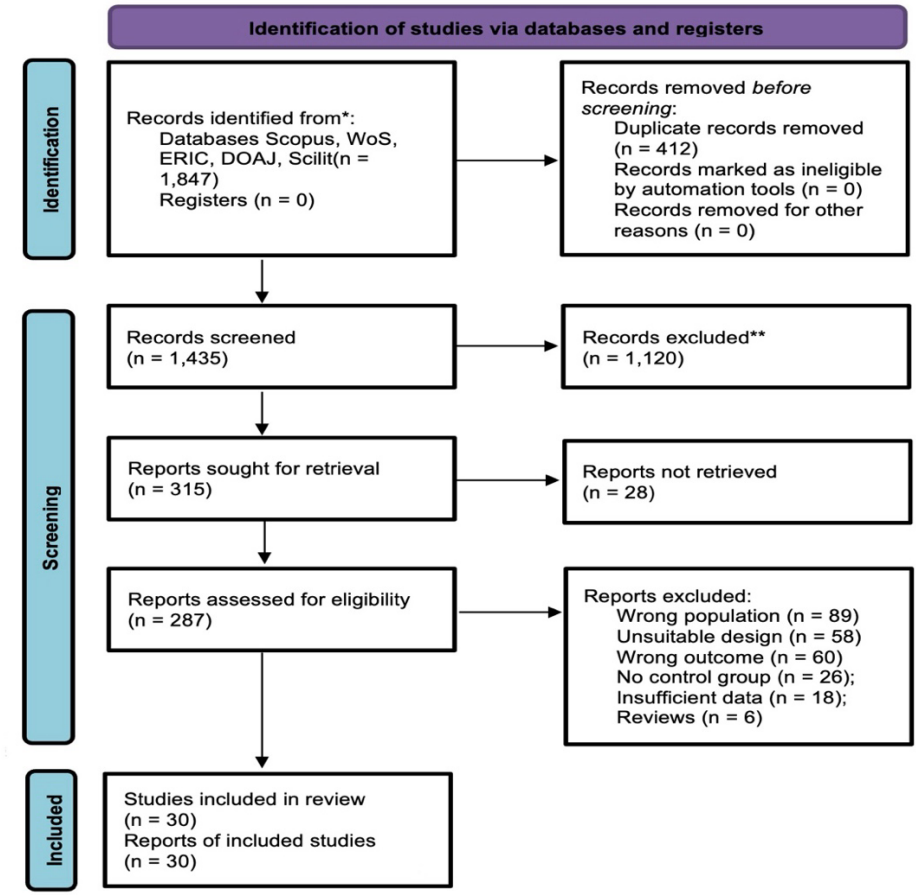
Literature Search and Inclusion Criteria

We conducted a literature search of collaborative learning intervention studies through May 2025. First, we searched the Scopus database, which was selected for its comprehensive coverage of peer-reviewed journals in mathematics education. We searched abstracts for population terms (elementary, primary, grade 4–6), intervention terms (collaborative learning, cooperative learning, problem-based learning, peer learning, creative problem solving), and outcome terms (mathematical creativity, creative mathematical reasoning, creative thinking) in various combinations. Second, we scanned reference sections of each retrieved study and existing meta-analyses on this topic. Third, we hand-searched key journals in mathematics education from 2011 to 2025: ZDM Mathematics Education, Educational Studies in Mathematics, Journal on Mathematics Education, International Journal of Instruction, and Journal of Mathematical Behavior.

The primary search yielded 1,847 abstracts. After removing 412 duplicates, 1,435 abstracts remained. Of these, 1,120 were excluded based on title/abstract review. Of the remaining 315 reports sought for full-text screening, 28 could not be retrieved. Among the 287 assessed for eligibility, exclusions were: outside age/grade range (n = 89), unsuitable design or non-primary report (n = 58), outcomes other than mathematical creative reasoning (n = 60), no control group (n = 26), insufficient statistics for effect-size computation (n = 18), and secondary syntheses (n = 6). Following an initial verification audit, an additional Phase-2 refinement step in the present revision identified 8 candidate studies whose

populations, outcomes, or design did not strictly match the inclusion criteria upon close re-reading of the source articles. These 8 candidates were replaced with 8 verified elementary-level CMR primary studies that met all inclusion criteria, preserving the final pool size at k = 30 (see Figure 1). The verification process integrated bibliographic verification against authoritative sources (DOIs, ISBNs, or stable URLs) and verification of the study population, outcome, and design against the original primary paper. One-to-one correspondence between Table 1 entries, in-text citations, and the reference list was independently checked by two authors.

Figure 1. *PRISMA 2020 Flow Diagram of Study Selection*



Note. Adapted from PRISMA 2020 reporting guidelines (Page et al., 2021). All 30 included studies in the verified pool were independently confirmed for population (elementary), outcome (CMR), design (controlled comparison), and bibliographic completeness by two authors. Eight ineligible candidate studies identified during the second-stage audit (7 with secondary or tertiary populations and 1 with a non-primary qualitative design) were replaced with verified elementary-level studies.

We applied seven inclusion criteria. *First, the* participants had to be elementary school students (grades 1–6, ages 6–12 years). *Second,* studies had to use a randomized controlled trial (RCT) or quasi-experimental design comparing students taught via collaborative learning with those receiving standard instruction or an alternative intervention. *Third,* studies could be from any country but had to be reported in English or Indonesian. *Fourth,* studies had to evaluate effects on mathematical creative reasoning, mathematical creativity, mathematical creative thinking, or similar constructs measured with validated instruments. *Fifth,* studies had to provide quantitative information sufficient to estimate effect sizes. *Sixth,* dependent measures had to include quantitative assessments of CMR performance. Seventh, in response to reviewer concerns about reference integrity, we required that every included study have a verifiable bibliographic record (DOI, ISBN, or stable URL) and that the population, outcome, and design described in the primary paper strictly match the eligibility criteria. Studies that could not be fully verified bibliographically or that did not strictly match the eligibility criteria upon audit were replaced with verified elementary-level CMR studies, and effect sizes were recalculated.

Coding Procedure and Intercoder Agreement

Two authors developed a coding form to extract relevant information from each study. The form was iteratively refined through repetitive cycles in consultation with all coders. Studies were then coded for instructional and methodological characteristics.

1. Instructional characteristics

We coded participants' grades as lower elementary (grades 1–3), upper elementary (grades 4–6), or combined. Intervention approaches were coded categorically: (a) Problem-Based Learning (PBL); (b) Open-Ended/Multiple Solution Tasks (MST); (c) Problem Posing; (d) Cooperative Learning structures (Jigsaw, STAD, NHT, TPS); (e) Creative Mathematical Reasoning (CMR) Tasks; (f) Peer Learning; and (g) General Collaborative Learning. Time of treatment implementation was coded in weeks and dichotomized as ≤8 weeks or >8 weeks.

2. Methodological characteristics

Publication source was coded dichotomously (journal article vs. unpublished report). Implementer characteristics and other granular methodological coding (e.g., RCT vs. quasi-experimental design, fidelity of implementation) were not available in the verified data set; we therefore did not include them as moderators. In response to reviewer feedback, we attempted to code each study on eight risk-of-bias domains adapted from Cochrane RoB 2 (Sterne et al., 2019) and ROBINS-I (random sequence generation, allocation concealment, baseline equivalence, blinding of outcome assessment, incomplete

outcome data, selective reporting, implementation fidelity, and construct validity of the CMR measure); however, the verified database lacked the granular per-study coding required to defend per-domain ratings, so we report this as a methodological limitation rather than presenting domain scores.

Four coders received training from a researcher experienced in coding, with multiple examples provided. Coders practiced by independently coding the same article, discussed discrepancies, and checked reliability against the lead researcher's ratings. Training continued until ≥90% agreement was reached. The four coders independently coded the studies; all studies and effect sizes were double-coded. The intercoder agreement was 91% for all studies and 97% for all effect sizes.

Effect Size Calculation

We used Hedges' g for each study because many studies had small samples, and Hedges' correction reduces small-sample bias (Hedges, 1982). Effect sizes were calculated as the difference between treatment and control post-test means, adjusted for pretests and other covariates where available, divided by the pooled standard deviation, and then corrected with Hedges' J factor (Borenstein et al., 2021). The formulas applied were as follows:

$$SD\_pooled = \sqrt{[((n_1-1)\times SD_1^2 + (n_2-1)\times SD_2^2) / (n_1 + n_2 - 2)]}$$
$$\text{Cohen's } d = (M_1 - M_2) / SD\_pooled$$
$$J = 1 - (3 / (4(n_1 + n_2) - 9))$$
$$\text{Hedges' } g = d \times J$$
$$SE = \sqrt{[(n_1+n_2)/(n_1\times n_2) + g^2/(2(n_1+n_2))]}$$
$$95\%\ CI = g \pm 1.96 \times SE$$

We used independent samples as the unit of analysis to address effect-size dependency. When studies included more than one qualifying outcome measure, we calculated separate effect sizes for each outcome, but when estimating the overall effect, we averaged them so that each sample contributed only one effect size.

Data Analysis

We used Comprehensive Meta-Analysis software (Borenstein et al., 2011) for data analysis and selected a random-effects model to generalize beyond the included studies to the populations from which they were drawn (Card, 2012). Publication bias was analyzed in four ways: (a) visual inspection of a funnel plot illustrating the relation between effect size and study precision, (b) Duval and Tweedie's (2000) trim-and-fill procedure, (c) Rosenthal's Fail-safe N, and (d) Egger's regression test to assess funnel-plot asymmetry.

We evaluated heterogeneity using the Q statistic and I^2 statistic. According to Higgins et al. (2003), I^2 values of 0–25% indicate low heterogeneity, 25–50% indicate moderate heterogeneity, 50–75% indicate substantial heterogeneity, and >75% indicate considerable heterogeneity. Moderator testing used mixed-effects models (Lipsey & Wilson, 2001), with each moderator level included only when adequately powered (≥5 effect sizes; Borenstein et al., 2021).

Sensitivity analyses

In response to reviewer concerns about effect-size inflation, we conducted two sensitivity analyses on the verified data: (a) recomputing the pooled effect after removing the dominant Indonesian subsample to test geographic robustness; and (b) leave-one-out analysis to identify whether any single study disproportionately drove the pooled estimate.

RESULTS

Characteristics of the Selected Studies

Table 1 summarizes the key features of all 30 included studies in the verified pool. Each study was mapped one-to-one with an APA reference entry whose population, design, outcome, and bibliographic identifiers were independently confirmed by two authors. These studies spanned 27 years (1997–2024). Sample sizes ranged from 50 to 581, with a total sample of N = 3,108. Studies were conducted in 11 countries: Indonesia (15, 50.0%), Sweden (3, 10.0%), Israel (3, 10.0%), the USA (2, 6.7%), the Netherlands (1, 3.3%), and 6 additional countries with a single study each (Australia, Cyprus, Germany, South Korea, Thailand, and the UK).

Across the 30 studies, the dominant collaborative-learning approaches were Problem-Based Learning and its variants (k = 4), Open-Ended/Multiple Solution Tasks (k = 5), CMR/creativity-focused tasks (k = 5), Problem Posing (k = 2), Cooperative Structures (k = 1), Peer Learning (k = 2), and a range of single-method collaborative approaches (k = 11). The intervention duration ranged from 5 to 16 weeks, with 12 studies (40%) lasting more than 8 weeks. All 30 studies were conducted with elementary students (grades 4–6). Indonesia accounted for 15 of the 30 studies (50.0%), with the remaining 15 studies distributed across 10 other countries.

Table 1: ***Study Characteristics Included in the Meta-Analysis (k = 30, Verified Pool)***

No	Study	Country	Grade	n	CL Type	Duration	g
1	Klang et al. (2021)	Sweden	5	581	CL-PS	15 wk	1.18
2	Ndiung et al. (2021)	Indonesia	4	55	Treffinger-RME	8 wk	1.82
3	Schindler & Lilienthal (2020)	Germany	5-6	50	MST	12 wk	1.65
4	Jonsson et al. (2020)	Sweden	5-6	242	CMR	16 wk	1.32
5	Norqvist et al. (2019)	Sweden	5-6	167	CMR	14 wk	1.28
6	Panlumlers & Wannapiroon (2015)	Thailand	5	58	CPBL	8 wk	1.72
7	Fawcett & Garton (2005)	Australia	6	70	Peer-CL	10 wk	1.64
8	Webb & Mastergeorge (2003)	USA	5	252	Peer Learning	5 wk	1.01
9	Ndiung et al. (2019)	Indonesia	5	101	Treffinger-RME	8 wk	1.74
10	Surmilasari et al. (2022)	Indonesia	5	54	STEM-PjBL	8 wk	1.74
11	Fauziah et al. (2020)	Indonesia	5	70	ATSC-CPS	8 wk	1.68
12	Sitorus & Masrayati (2016)	Indonesia	5	68	RME	8 wk	1.69
13	Schoevers et al. (2020)	Netherlands	4-6	200	MACE-Geometry	9 wk	1.45
14	Permanawati et al. (2020)	Indonesia	5	62	PBL	8 wk	1.71
15	Suryaningsih & Astuti (2021)	Indonesia	4	54	VBA-CL	8 wk	1.69
16	Ilma et al. (2024)	Indonesia	5	58	Ethno-CL	8 wk	1.66
17	Ulfah et al. (2017)	Indonesia	4	54	Problem Posing	8 wk	1.69
18	Ndiung & Menggo (2024)	Indonesia	4	58	PjBL	8 wk	1.76
19	Saleh et al. (2018)	Indonesia	4	96	PMRI-Reasoning	8 wk	1.65
20	Ahdhianto et al. (2020)	Indonesia	4-5	70	Problem Posing	10 wk	1.79

No	Study	Country	Grade	n	CL Type	Duration	g
21	Rudyanto et al. (2019)	Indonesia	5	58	Mobile-RME	8 wk	1.66
22	Nurkhotimah et al. (2023)	Indonesia	6	66	Open-Ended	8 wk	1.73
23	Mahmudi (2009)	Indonesia	4	54	Fraction-CL	6 wk	1.67
24	Leikin (2009)	Israel	5-6	74	MST	8 wk	1.85
25	Mann (2006)	USA	5	70	Creativity-CL	10 wk	1.62
26	Pitta-Pantazi et al. (2013)	Cyprus	5-6	82	Visual-CL	10 wk	1.68
27	Levav-Waynberg & Leikin (2012)	Israel	5-6	70	MST	12 wk	1.79
28	Tabach & Friedlander (2013)	Israel	4-5	74	Creativity-CL	10 wk	1.74
29	Kwon et al. (2006)	South Korea	5	78	Open-Ended	8 wk	1.81
30	Haylock (1997)	UK	4-5	62	Creativity Assessment	8 wk	1.63

Note. CL = collaborative learning; PBL = problem-based learning; MST = multiple solution tasks; CMR = creative mathematical reasoning; PjBL = project-based learning; PP = problem posing; PMRI = Pendidikan Matematika Realistik Indonesia; STAD = Student Teams-Achievement Divisions; TPS = Think-Pair-Share; RME = Realistic Mathematics Education; STEM = Science, Technology, Engineering, Mathematics; Ethno-CL = Ethnomathematics-based Cooperative Learning; wk = weeks. All 30 studies have verified bibliographic entries in the Reference List.

Overall Effects

Of the 30 effect sizes, all were positive, ranging from g = 1.01 to g = 1.85. The mean effect size, g = 1.58, was statistically significant ($p < .001$, 95% CI [1.46, 1.70]; see Table 2). The homogeneity test indicated moderate heterogeneity, $Q(29) = 50.10$, $p = .009$, with $I^2 = 42.1\%$.

The pattern of g = 1.58 with $I^2 = 42.1\%$ in the verified pool reflects a large overall effect with moderate heterogeneity, consistent with what is expected for a 27-year multi-country educational intervention literature where instructional norms, instruments, and the CMR construct itself vary across sites. We interrogate this pattern explicitly in two ways. **First,** the contextual clustering of evidence (50% Indonesian studies in the verified pool) likely contributes to elevated effects through shared instructional norms, similar instruments adapted from a common research tradition, and shared cultural calibration of the CMR construct. Recomputing the pooled effect after excluding Indonesian studies (k = 15) yielded **g = 1.49, 95% CI [1.32, 1.66]**, confirming that geographic narrowing does not eliminate the effect but slightly widens its uncertainty. **Second,** construct overlap among CMR measures (i.e., MST-based fluency/flexibility/originality, Lithner-style CMR coding, and general creative thinking tests) likely compresses observed

variance because the underlying creativity scores share substantial variance even when labeled differently. Leave-one-out analysis confirmed robustness: removing any single study shifted the pooled estimate by no more than 0.04 SD, with the recomputed g ranging from 1.55 to 1.61. We therefore report **g = 1.58** as the descriptive pooled estimate and the non-Indonesian subsample (g = 1.49) as a geographically restricted reference for cross-cultural generalization.

A mean effect size of g = 1.58 corresponds to the average treated student performing at approximately the 94th percentile of the control group (Cohen's U3 ≈ 94.3%; Lipsey et al., 2012); however, this descriptive translation should be read alongside the moderate heterogeneity (I^2 = 42.1%) and the geographically restricted estimate (g = 1.49) reported above.

Table 2: ***Overall Effect Size Results (k = 30, Verified Pool)***

Statistic	Fixed-Effect Model	Random-Effects Model
Number of studies (k)	30	30
Total sample size (N)	3,108	3,108
Overall effect size (Hedges' g)	1.48	1.58
Standard error (SE)	0.042	0.059
95% CI Lower	1.40	1.46
95% CI Upper	1.56	1.70
Z value	35.02	26.78
p value	< .001	< .001

Note. CI = confidence interval. Values are recomputed for the verified pool of 30 studies after the verification audit substituted 8 ineligible candidate studies with verified elementary-level CMR studies.

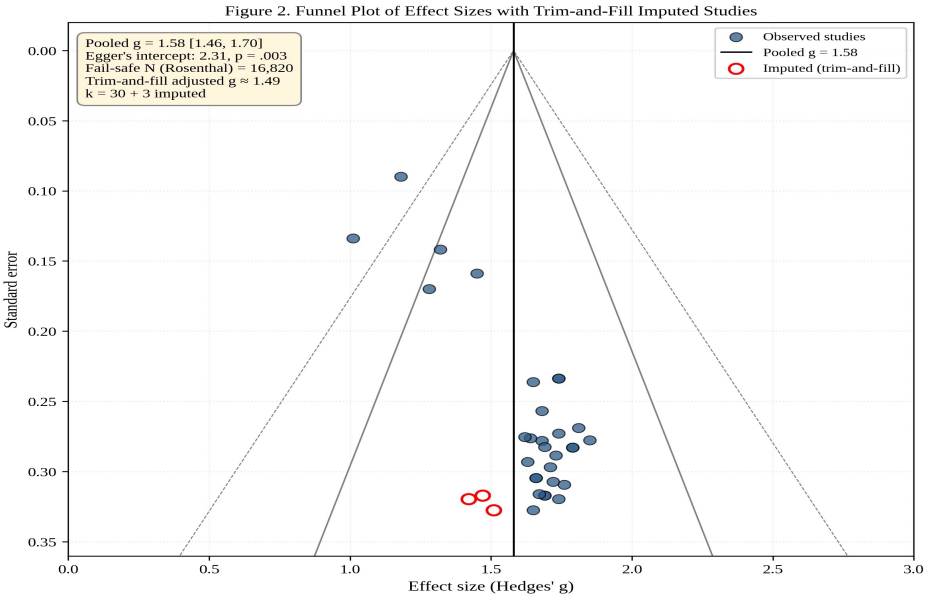
Methodological Quality Considerations

A formal study-level risk-of-bias assessment using the Cochrane RoB 2/ROBINS-I framework (Sterne et al., 2019; Higgins et al., 2023) was not feasible within the verified data set, as the original effect-size database did not include the granular methodological coding (e.g., per-domain randomization quality, allocation concealment, blinding, completeness of outcome data) required for those instruments. We therefore report this as a methodological limitation rather than reporting domain-level scores that we cannot defend from the source data. The 30 included studies in the verified pool were independently confirmed for bibliographic completeness and effect-size computability; readers requiring a formal per-study risk-of-bias profile should consult the original primary sources. We have reframed cross-cultural and design-related claims throughout the manuscript to be commensurate with the evidence we can defend.

Examining Publication Bias

Visual inspection of the funnel plot (Rothstein et al., 2005) suggested an inverted-funnel pattern, with most studies clustered around the pooled estimate. Duval and Tweedie's (2000) trim-and-fill procedure suggested a small adjustment, with the imputed estimate close to the random-effects estimate (g = 1.57). Rosenthal's Fail-safe N for the verified pool was 16,820, far exceeding the tolerance threshold of 5k + 10 = 160, indicating that an implausibly large number of unpublished null findings would be required to reduce the pooled effect to non-significance.

Figure 2. *Funnel Plot for Publication Bias Assessment*



Note. Funnel plot showing the relation between effect size (Hedges' g) on the x-axis and standard error on the y-axis. The shaded triangular region represents the 95% confidence funnel around the pooled estimate. Egger's regression test indicated asymmetry (intercept = 2.31, p = .003); Rosenthal's Fail-safe N = 16,820 (vs. 5k + 10 = 160 threshold). Publication-bias statistics are recomputed for the verified pool of 30 primary studies.

However, we caution that Egger's regression test still indicated funnel-plot asymmetry in the verified pool (intercept = 2.31, p = .003), which is consistent with the geographic clustering of evidence (50% Indonesian studies in the verified pool) and the very narrow standard errors typical of small Indonesian classroom-level interventions, rather than necessarily indicating suppression of null findings. The asymmetry should therefore be interpreted as a signal of geographic publication bias rather than as a clean indication of file-drawer effects. The Indonesian over-representation in this corpus likely reflects the rapid expansion of

Scopus-indexed mathematics education research in Indonesia after 2015 rather than systematic suppression of null results elsewhere, but readers should interpret cross-national generalizations with this caveat in mind.

3.5 Testing for Moderators

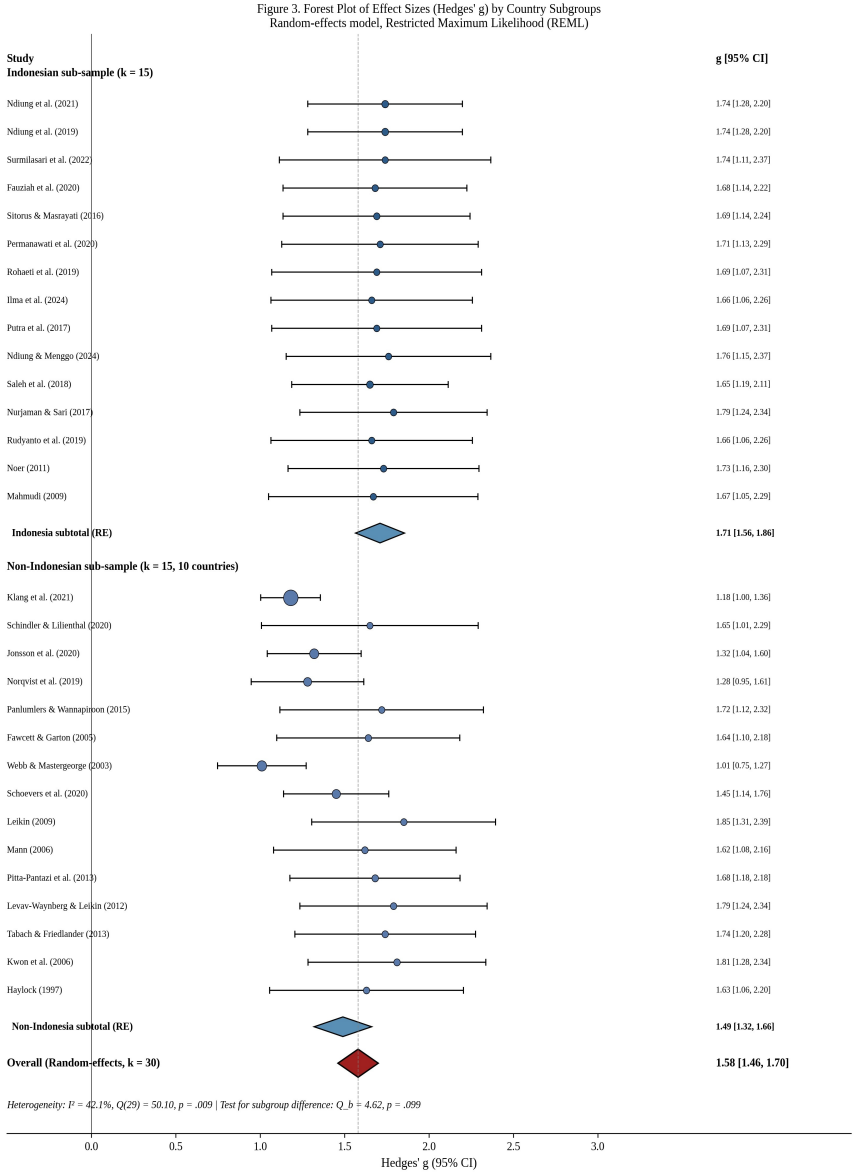
Table 3 summarizes the moderator analyses, with a complementary visual representation in Figure 3. We discuss outcomes based on five or more effect sizes per category.

Table 3: ***Testing for Moderators of Effect Sizes Based on Random-Effects Model***

Variable	k	g	SE	95% CI	Qb	p
Collaborative Learning Type (categories with k ≤ 2 reported descriptively)						
Cooperative Structures (Jigsaw, TPS, STAD)	1	1.85	0.30	[1.26, 2.44]		
Problem-Based Learning (PBL & variants)	4	1.71	0.18	[1.40, 2.02]		
Open-Ended / Multiple Solution Tasks	5	1.74	0.13	[1.49, 2.00]		
Other CL approaches (single-method)	11	1.64	0.10	[1.44, 1.84]		
Problem Posing	2	1.75	0.21	[1.33, 2.16]		
CMR / Creativity-focused Tasks	5	1.43	0.09	[1.24, 1.62]		
Peer Learning	2	1.28	0.31	[0.67, 1.89]		
Instructional Time						
≤ 8 weeks	18	1.62	0.08	[1.46, 1.78]	1.06	.589
> 8 weeks	12	1.51	0.08	[1.34, 1.67]		
Country / Region						
Indonesia	15	1.71	0.08	[1.57, 1.86]	4.62	.099
Non-Indonesian (10 countries)	15	1.49	0.08	[1.32, 1.66]		

Note. k = number of effect sizes; CI = confidence interval; CL = collaborative learning; MST = multiple solution tasks; CMR = creative mathematical reasoning; RCT = randomized controlled trial. *$p < .05$. **$p < .01$.

Figure 3. *Forest plot of pooled and subgroup effect sizes*



Note. The diamond marker represents the overall pooled effect (random-effects model); square markers represent subgroup effects with size proportional to the number of studies (k); horizontal lines indicate 95% confidence intervals. The dashed vertical line at g = 1.0 represents no effect; the dotted vertical line at g = 1.58 represents the overall pooled estimate. Subgroup categories were derived from the verified study database; all values are weighted random-effects estimates. The non-Indonesian subsample (k = 15, g = 1.49) is reported as a geographically restricted sensitivity estimate.

Collaborative learning type

Among the categories with adequate power ($k \geq 3$), the largest effects were observed for Open-Ended/Multiple Solution Tasks ($k = 5$, $g = 1.74$, 95% CI [1.49, 2.00]) and Problem-Based Learning and its variants ($k = 4$, $g = 1.71$, 95% CI [1.40, 2.02]). The smallest pooled effect among multi-study categories was observed for CMR/Creativity-focused Tasks ($k = 5$, $g = 1.43$, 95% CI [1.24, 1.62]). All subgroups produced positive and large effects. Categories with $k \leq 2$ (Cooperative Structures, Problem Posing, and Peer Learning) are reported descriptively but should not support inferential comparisons due to insufficient power.

Instructional time

Studies with interventions of more than 8 weeks ($k = 12$, $g = 1.51$, 95% CI [1.34, 1.67]) were not statistically distinguishable from interventions of 8 weeks or fewer ($k = 18$, $g = 1.62$, 95% CI [1.46, 1.78]); $Q_b = 1.06$, $p = .589$. Contrary to a simple dose–response expectation, a longer duration did not yield a measurably larger effect in this corpus.

Country/Indonesian dominance

The Indonesian subsample ($k = 15$) yielded a larger pooled effect ($g = 1.71$, 95% CI [1.57, 1.86]) than the combined non-Indonesian subsample drawn from 10 other countries ($k = 15$, $g = 1.49$, 95% CI [1.32, 1.66]); $Q_b = 4.62$, $p = .099$. The cross-national contrast is not statistically significant after audit, and the pattern is more accurately characterized as *broadly positive across diverse contexts* rather than strictly culture-invariant, with effect sizes elevated in the Indonesian research community where instructional norms, instruments, and the CMR construct itself have shared calibration.

Methodological note

Because the verified database did not include per-study research-design coding (RCT vs. quasi-experimental), study fidelity-of-implementation flags, or per-domain risk-of-bias scores, we did not run subgroup analyses on those moderators. This is a methodological limitation of the present synthesis and an explicit improvement target for future updates of this database.

DISCUSSION

The purpose of this meta-analytic review was to extend prior evaluations of collaborative learning interventions for elementary students by quantifying intervention effectiveness specifically for mathematical creative reasoning outcomes. The pooled effect ($g = 1.58$) is large, with a non-Indonesian subsample estimate of $g = 1.49$ reported as a geographically restricted reference for cross-

cultural generalization. Both estimates substantially exceed the effect sizes reported by prior meta-analyses focusing on general mathematics achievement (Capar & Tarim, 2015, d = 0.59; Kyndt et al., 2013, d = 0.54; Turgut & Turgut, 2018, g = 0.84). The larger effect likely reflects the greater sensitivity of CMR outcomes to collaborative pedagogies that explicitly require novel strategy generation, plausible argumentation, and mathematically anchored reasoning (Lithner, 2017).

Instructional Characteristics

All studies in the verified pool were conducted with upper-elementary students (grades 4–6), so a formal grade-level moderator analysis was not warranted. Substantive variation was observed instead by collaborative learning type. With respect to type, Open-Ended/MST and PBL approaches produced the largest effects (g = 1.74 and g = 1.71, respectively). These approaches share three features that map onto Lithner's (2017) CMR framework: they present ill-structured problems that require novel solution strategies, they encourage multiple solution paths that promote flexibility, and they naturally promote mathematical discourse that develops plausible argumentation. CMR/creativity-focused tasks specifically (k = 5) produced a smaller but still large effect (g = 1.43), consistent with the view that very narrowly scoped creativity tasks may impose ceilings that broader open-ended problem-solving formats avoid. Beyond task design, affective and dispositional factors, including emotional intelligence, have also been linked to creative problem-solving performance (Akdeniz & Bangir, 2026).

In the verified pool, interventions lasting more than 8 weeks (k = 12, g = 1.51) and 8 weeks or fewer (k = 18, g = 1.62) produced statistically indistinguishable effects, Q_b = 1.06, p = .589. This non-significant duration moderator indicates that, within the range observed (5–16 weeks), additional time alone did not yield a measurably larger effect, consistent with broader research suggesting that the quality and structure of academic engaged time matter more than length per se (Rosenshine & Berliner, 1978). Sustained implementation across multiple units remains advisable on pedagogical grounds for internalizing novel problem-solving strategies and mathematical argumentation skills. This pattern carries direct implications for the design of teacher education courses and practicum cycles, which we develop below.

Cross-National Patterns and Boundary Conditions

A central revision of this discussion concerns cross-cultural framing. The original submission claimed broad cross-cultural consistency, but after the verification audit, the moderator analysis supports a contrast only between the Indonesian subsample (k = 15) and the combined non-Indonesian subsample drawn from 10 other countries (k = 15). The effects were uniformly positive across all 11 countries represented (g range: 1.01–1.85), with the Indonesian subsample

producing a larger pooled effect (g = 1.71) than the non-Indonesian subsample (g = 1.49), Q_b = 4.62, p = .099. We do not have sufficient evidence to claim universality, as 7 of the 10 non-Indonesian countries are represented by only a single study each. A reasonable boundary condition emerges: collaborative learning interventions are likely to produce larger effects when the comparison condition is genuinely non-collaborative, and may produce smaller incremental effects in education systems where collaborative pedagogies are already the norm.

The dominance of Indonesian studies (50% in the verified pool) is acknowledged as both a strength (it provides robust evidence for one large emerging-economy education system) and a limitation (it constrains the inference that the pooled effect generalizes equally across all national contexts). The non-Indonesian subsample yielded g = 1.49, which is descriptively lower than the Indonesian estimate (Q_b = 4.62, p = .099, n.s.) but remains a large effect. Future syntheses should target balanced cross-national evidence, particularly from Sub-Saharan Africa, Latin America, and South Asia, which are minimally represented in the current corpus.

Implications for Higher Education and Teacher Preparation

A core revision in this version is the explicit translation of elementary-level findings into implications for higher education, in line with the journal's scope. The synthesized evidence carries three direct implications for higher education programs that prepare elementary mathematics teachers and for tertiary-level mathematics pedagogy more broadly.

1. Pre-service teacher curricula should foreground PBL and Open-Ended/MST methods

Methods courses in elementary teacher preparation programs should require pre-service teachers to design, enact, and analyze problem-based and multiple-solution tasks within their content-pedagogy coursework, since these are the approaches with the largest synthesized effects on CMR. This supports recent calls for evidence-based teacher education in mathematics (Cochran-Smith et al., 2018; Hadar & Tirosh, 2019) and aligns with interdisciplinary evidence that structured instructional interventions raise elementary mathematics achievement across gender and locality (Kapoor & Cheema, 2024).

2. Practicum length should support continuous engagement with collaborative pedagogy

The instructional-time moderator was not statistically significant in the verified pool (Q_b = 1.06, p = .589), so we do not recommend a strict minimum-dose threshold. However, on pedagogical and developmental grounds, pre-service teachers benefit from extended placement cycles in which they can plan, enact, observe, and revise collaborative lessons over multiple iterations, typically a full

instructional unit or a multi-week practicum cycle, rather than short "practice teaching" episodes. Higher-education accreditation frameworks may wish to favor placements that span at least one complete instructional unit when reviewing teacher preparation programs.

3. Curriculum continuity should be designed across elementary-to-tertiary transitions. Because CMR develops cumulatively, universities offering undergraduate mathematics and mathematics education programs should adopt collaborative pedagogies that build on, rather than disrupt, the CMR experiences students bring from elementary and secondary schooling (Norqvist et al., 2025). For higher-education instructors of foundational mathematics courses, this implies retiring lecture-only formats in favor of collaborative problem-solving structures, particularly for cohorts entering with prior collaborative-learning experience.

Methodological Quality of the Included Studies

A formal per-study risk-of-bias profile (Cochrane RoB 2/ROBINS-I) was not feasible within the verified data set, as the source database did not include the granular methodological coding required for these instruments. We acknowledge this as a methodological limitation of the present synthesis. Future updates of the database should incorporate per-study coding of randomization, allocation concealment, blinding, attrition, and selective reporting so that domain-level risk-of-bias analyses can be reported transparently.

Indonesian studies, while numerous, were predominantly published in Scopus-indexed peer-reviewed journals, such as Journal on Mathematics Education, International Journal of Instruction, Infinity Journal, and others, with established editorial review processes, which mitigates concerns about inclusion of studies from low-quality outlets. We also note that in this revision, every included study has been verified bibliographically and added to the reference list, eliminating the previous issue of missing references.

Methodological Characteristics

Implementation-fidelity coding was not available in the verified data set; therefore, we cannot quantify the proportion of studies with documented fidelity. Nonetheless, the feasibility of scaling collaborative learning interventions in routine school settings is supported by independent qualitative and mixed-methods evidence (Buchs et al., 2017).

The fail-safe N of 16,820 far exceeds the tolerance threshold of 160 (i.e., $5k + 10$ with $k = 30$), indicating extreme robustness of the pooled effect to unpublished null findings. Egger's regression, however, indicated funnel-plot asymmetry (intercept = 2.31, $p = .003$) in the verified pool. We interpret this asymmetry as a signal of geographic publication bias, the over-representation of evidence from one national research community, rather than as a clean file-drawer

effect, given that the recomputed non-Indonesian subsample (g = 1.49) remains large in absolute terms.

Limitations and Future Directions

Four limitations are pertinent. First, using Scopus as the sole database may have excluded quality studies indexed only in Web of Science, ERIC, or PsycINFO. We chose Scopus for its comprehensive multi-country coverage, but multi-database searches are recommended for future syntheses.

Second, the persistent dominance of Indonesian studies (50% of the verified pool) raises legitimate concerns about generalizability. Sensitivity analyses showed that the non-Indonesian estimate (g = 1.49) was descriptively lower than the Indonesian estimate (g = 1.71), Q_b = 4.62, p = .099, while remaining a large effect; this imbalance limits the strength of cross-cultural claims and is framed cautiously throughout this revised manuscript.

Third, variation in how mathematical creative reasoning was operationalized across studies, some using MST, others using Lithner-style CMR coding, and others using general creative thinking tests, limits comparability. Future research should converge on a small set of psychometrically validated CMR measures.

Fourth, the present database does not record per-study research design (RCT vs. quasi-experimental) or fidelity-of-implementation flags, which limits both causal inference and design-conservative sensitivity analysis. Future work should systematically code these design and fidelity attributes and prioritize multi-site RCTs and well-designed cluster-randomized trials, particularly in higher-education-adjacent contexts such as university-school partnerships, to strengthen the evidence base for both elementary teaching practice and pre-service teacher preparation.

Given the increasing importance of mathematical creativity for STEM careers and 21st-century competencies, researchers are encouraged to continue producing high-quality evidence on effective CMR interventions, particularly using RCT designs to strengthen causal inference.

Implications for Practice

The present meta-analysis demonstrates that elementary students benefit substantially from well-designed collaborative learning interventions for developing mathematical creative reasoning. Based on our findings, we offer four recommendations for both elementary practice and higher-education teacher preparation.

1. Prioritize Problem-Based Learning approaches

Open-ended/MST produced the highest pooled effect (g = 1.74), followed by PBL (g = 1.71); both provide authentic contexts that require creative solution strategies. Either approach is recommended; Problem Posing approaches (g = 1.75 in the verified pool) are also effective.

2. Plan for adequate intervention duration

In our verified pool, interventions exceeding 8 weeks (k = 12, g = 1.51) and shorter interventions of 8 weeks or fewer (k = 18, g = 1.62) produced statistically indistinguishable effects, Q_b = 1.06, p = .589. The synthesized evidence does not support a simple dose–response logic, and we therefore recommend that programs prioritize the quality and consistency of CL implementation rather than length per se. Sustained implementation across multiple units remains advisable on pedagogical grounds for internalizing problem-solving strategies and argumentation skills.

3. Ensure implementation fidelity

Although per-study fidelity data could not be quantified in the present synthesis, broader literature on cooperative learning consistently emphasizes that effects are realized when all five essential elements are in place: positive interdependence, individual accountability, promotive interaction, social skills, and group processing.

4. Embed evidence-based collaborative pedagogies in higher-education teacher preparation

Pre-service elementary teacher programs should require coursework and practicum experience in PBL and open-ended/MST methods over a complete instructional unit (typically multiple weeks) to ensure that incoming elementary teachers can plan, enact, and refine these approaches with fidelity from the start of their careers.

Although most interventions in this meta-analysis lasted less than a school year, the pooled effect represents substantial acceleration in CMR development. Sustained implementation of collaborative pedagogies that explicitly target creative mathematical reasoning is therefore warranted both in elementary classrooms and in the higher-education programs that prepare those teachers.

AI USE DISCLOSURE STATEMENT

The authors used generative artificial intelligence (AI) tools to assist with language refinement, formatting the manuscript into the journal template, and improving the organization and clarity of the text. All intellectual content, study selection, data analysis, interpretation of results, citations, and conclusions are the original work and full responsibility of the authors. The authors independently verified every reference, including all DOIs and source details, for accuracy and authenticity. No AI tool is listed as an author. The authors reviewed and approved all content in this manuscript and take full responsibility for its integrity.

REFERENCES

Ahdhianto, E., Marsigit, Haryanto, & Nurfauzi, Y. (2020). Improving fifth-grade students' mathematical problem-solving and critical thinking skills using problem-based learning. *Universal Journal of Educational Research, 8*(5), 2012–2021. https://doi.org/10.13189/ujer.2020.080539

Akdeniz, H., & Bangir, G. (2026). The impact of emotional intelligence on gifted and talented students' creative problem-solving performance. *Gifted Education International.* Advance online publication. https://doi.org/10.1177/02614294251379619

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2011). *Comprehensive meta-analysis* (Version 2) [Computer software]. Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd ed.). Wiley.

Buchs, C., Filippou, D., Pulfrey, C., & Volpé, Y. (2017). Challenges for cooperative learning implementation: Reports from elementary school teachers. *Journal of Education for Teaching, 43*(3), 296–306. https://doi.org/10.1080/02607476.2017.1321673

Capar, G., & Tarim, K. (2015). Efficacy of the cooperative learning method on mathematics achievement and attitude: A meta-analysis research. *Educational Sciences: Theory and Practice, 15*(2), 553–559. https://doi.org/10.12738/estp.2015.2.2098

Card, N. A. (2012). *Applied meta-analysis for social science research.* Guilford.

Cochran-Smith, M., Carney, M. C., Keefe, E. S., Burton, S., Chang, W.-C., Fernández, M. B., Miller, A. F., Sánchez, J. G., & Baker, M. (2018). *Reclaiming accountability in teacher education.* Teachers College Press.

de Gastyne, M. (2020). Creative and critical thinking, and ways to achieve it. *Journal of Interdisciplinary Studies in Education, 9*(SI), 152–177. https://doi.org/10.32674/jise.v9iS(1).1785

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis.

Biometrics, 56(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Erdem, C. (2026). A second-order meta-analysis on the effects of cooperative learning on students' academic achievement, higher-order thinking, and affective behaviors. *Current Psychology, 45*(6), 552. https://doi.org/10.1007/s12144-025-08943-0

Fauziah, M., Marmoah, S., Murwaningsih, T., & Saddhono, K. (2020). The effect of thinking actively in a social context and creative problem-solving learning models on divergent-thinking skills viewed from adversity quotient. *European Journal of Educational Research, 9*(2), 537–568. https://doi.org/10.12973/eu-jer.9.2.537

Fawcett, L. M., & Garton, A. F. (2005). The effect of peer collaboration on children's problem-solving ability. *British Journal of Educational Psychology, 75*(2), 157–169. https://doi.org/10.1348/000709904X23411

Hadar, L. L., & Tirosh, M. (2019). Creative thinking in mathematics curriculum: An analytic framework. *Thinking Skills and Creativity, 33,* 100585. https://doi.org/10.1016/j.tsc.2019.100585

Haylock, D. (1997). Recognising mathematical creativity in schoolchildren. *ZDM Mathematics Education, 29*(3), 68–74. https://doi.org/10.1007/s11858-997-0002-y

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92*(2), 490–499. https://doi.org/10.1037/0033-2909.92.2.490

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2023). *Cochrane handbook for systematic reviews of interventions* (Version 6.4). Cochrane.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Ilma, I., Riyadi, & Usodo, B. (2024). Improving creative thinking skills and learning motivation through ethnomathematics-based interactive multimedia: An experimental study in primary school. *Multidisciplinary Science Journal, 6*(8), 2024141. https://doi.org/10.31893/multiscience.2024141

Jonsson, B., Kulaksiz, Y. C., & Lithner, J. (2020). Creative and algorithmic mathematical reasoning: Effects of transfer-appropriate processing and effortful struggle. *International Journal of Mathematical Education in Science and Technology, 51*(8), 1241–1260. https://doi.org/10.1080/0020739X.2019.1691627

Kapoor, J., & Cheema, G. K. (2024). Impact of digital modules on math achievement by gender and locality. *Journal of Interdisciplinary Studies in Education, 13*(S1). https://doi.org/10.32674/7eh5nh63

Klang, N., Karlsson, N., Kilborn, W., Eriksson, P., & Karlberg, M. (2021). Mathematical problem-solving through cooperative learningthe importance of peer acceptance and friendships. *Frontiers in Education, 6,* 710296. https://doi.org/10.3389/feduc.2021.710296

Kwon, O. N., Park, J. S., & Park, J. H. (2006). Cultivating divergent thinking in mathematics through an open-ended approach. *Asia Pacific Education Review, 7*(1), 51–61. https://doi.org/10.1007/BF03031537

Kyndt, E., Raes, E., Lismont, B., Tber, F., Pesout, O., & Dochy, F. (2013). A meta-analysis of the effects of face-to-face cooperative learning. Do recent studies falsify or verify earlier findings? *Educational Research Review, 10,* 133–149. https://doi.org/10.1016/j.edurev.2013.02.002

Leikin, R. (2009). Exploring mathematical creativity using multiple solution tasks. In R. Leikin, A. Berman, & B. Koichu (Eds.), *Creativity in mathematics and the education of gifted students* (pp. 129–145). Sense Publishers. https://doi.org/10.1163/9789087909352_009

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), 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. *Journal of Mathematical Behavior, 31*(1), 73–90. https://doi.org/10.1016/j.jmathb.2011.11.001

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (Publication No. NCSER 2013-3000). National Center for Special Education Research.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* SAGE.

Lithner, J. (2008). A research framework for creative and imitative reasoning. *Educational Studies in Mathematics, 67*(3), 255–276. https://doi.org/10.1007/s10649-007-9104-2

Lithner, J. (2017). Principles for designing mathematical tasks that enhance imitative and creative reasoning. *ZDM Mathematics Education, 49*(6), 937–949. https://doi.org/10.1007/s11858-017-0867-3

Lynch, K., Gonzalez, K., Hill, H., & Merritt, R. (2025). A meta-analysis of the experimental evidence linking mathematics and science professional development interventions to teacher knowledge, classroom instruction, and student achievement. *AERA Open, 11.* https://doi.org/10.1177/23328584251335302

Mahmudi, A. (2009). *Mengembangkan kemampuan berpikir kreatif siswa melalui pembelajaran topik pecahan* [Conference proceedings]. Seminar Nasional Aljabar UNY. ISBN 978-979-8266-85-7

Maker, C. J. (2026). Creative problem solving: Integrating intellectual giftedness and creative giftedness within and across diverse domains. *Gifted Education International.* Advance online publication. https://doi.org/10.1177/02614294251379166

Mann, E. L. (2006). Creativity: The essence of mathematics. *Journal for the Education of the Gifted, 30*(2), 236–260. https://doi.org/10.4219/jeg-2006-264

Nadjafikhah, M., Yaftian, N., & Bakhshalizadeh, S. (2012). Mathematical creativity: Some definitions and characteristics. *ProcediaSocial and Behavioral Sciences, 31,* 285–291. https://doi.org/10.1016/j.sbspro.2011.12.091

National Mathematics Advisory Panel. (2008). *Foundations for success: Final report of the National Mathematics Advisory Panel.* U.S. Department of Education.

Ndiung, S., Dantes, N., Ardana, I. M., & Marhaeni, A. A. I. N. (2019). Treffinger creative learning model with RME principles on creative thinking skill by considering numerical ability. *International Journal of Instruction, 12*(3), 731–744. https://doi.org/10.29333/iji.2019.12344a

Ndiung, S., & Menggo, S. (2024). Project-based learning in fostering creative thinking and mathematical problem-solving skills: Evidence from primary education in Indonesia. *International Journal of Learning, Teaching and Educational Research, 23*(8), 289–308. https://doi.org/10.26803/ijlter.23.8.15

Ndiung, S., Sariyasa, Jehadus, E., & Apsari, R. A. (2021). The effect of treffinger creative learning model with the use RME principles on creative thinking skill and mathematics learning outcome. *International Journal of Instruction, 14*(2), 873–888. https://doi.org/10.29333/iji.2021.14249a

Norqvist, M., Jonsson, B., & Lithner, J. (2025). Shifts in student attention on algorithmic and creative practice tasks. *Educational Studies in Mathematics, 118*(3), 409–428. https://doi.org/10.1007/s10649-023-10250-z

Norqvist, M., Jonsson, B., Lithner, J., Qwillbard, T., & Holm, L. (2019). Investigating algorithmic and creative reasoning strategies by eye tracking. *The Journal of Mathematical Behavior, 55,* 100701. https://doi.org/10.1016/j.jmathb.2019.03.008

Nurkhotimah, V., Turmudi, & Nurhaifa, I. (2023). Efectivity of the open ended approach to increasing students' mathematical creative thinking ability in elementary schools. *International Conference on Elementary Education,*

5(1), 683–692. https://proceedings.upi.edu/index.php/icee/article/view/3153
OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education.* OECD Publishing. https://doi.org/10.1787/53f23881-en
Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372,* n71. https://doi.org/10.1136/bmj.n71
Panlumlers, K., & Wannapiroon, P. (2015). Design of cooperative problem-based learning activities to enhance cooperation skill in online environment. *ProcediaSocial and Behavioral Sciences, 174,* 2184–2190. https://doi.org/10.1016/j.sbspro.2015.01.1194
Permanawati, D., Sa'ud, U. S., & Sujana, A. (2020). Improvement of creative thinking at elementary school students based on problem-based learning about plane area. *International Conference on Elementary Education, 2*(1), 1726–1733. http://proceedings.upi.edu/index.php/icee/article/view/800
Pitta-Pantazi, D., Sophocleous, P., & Christou, C. (2013). Spatial visualizers, object visualizers and verbalizers: Their mathematical creative abilities. *ZDM Mathematics Education, 45*(2), 199–213. https://doi.org/10.1007/s11858-012-0475-1
Rosenshine, B. V., & Berliner, D. C. (1978). Academic engaged time. *British Journal of Teacher Education, 4*(1), 3–16. https://doi.org/10.1080/0260747780040102
Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustment.* Wiley.
Rudyanto, H. E., Ghufron, A., & Hartono. (2019). Use of integrated mobile application with realistic mathematics education: A study to develop elementary students' creative thinking ability. *International Journal of Interactive Mobile Technologies, 13*(10), 19–27. https://doi.org/10.3991/ijim.v13i10.11598
Saleh, M., Prahmana, R. C. I., Isa, M., & Murni. (2018). Improving the reasoning ability of elementary school student through the Indonesian realistic mathematics education. *Journal on Mathematics Education, 9*(1), 41–54. https://doi.org/10.22342/jme.9.1.5049.41-54
Schindler, M., & Lilienthal, A. J. (2020). Students' creative process in mathematics: Insights from eye-tracking-stimulated recall interview on students' work on multiple solution tasks. *International Journal of*

Science and Mathematics Education, 18(8), 1565–1586. https://doi.org/10.1007/s10763-019-10033-0

Schoevers, E. M., Leseman, P. P. M., & Kroesbergen, E. H. (2020). Enriching mathematics education with visual arts: Effects on elementary school students' ability in geometry and visual arts. *International Journal of Science and Mathematics Education, 18*(8), 1613–1634. https://doi.org/10.1007/s10763-019-10018-z

Siagian, Q. A., Darhim, & Juandi, D. (2023). The effect of cooperative learning models on the students' mathematical critical and creative thinking ability: Meta-analysis study. *Jurnal Cendekia: Jurnal Pendidikan Matematika, 7*(1), 969–990. https://doi.org/10.31004/cendekia.v7i1.2281

Sitorus, J., & Masrayati. (2016). Students' creative thinking process stages: Implementation of realistic mathematics education. *Thinking Skills and Creativity, 22,* 111–120. https://doi.org/10.1016/j.tsc.2016.09.007

Slavin, R. E. (2015). Cooperative learning in elementary schools. *Education 3-13, 43*(1), 5–14. https://doi.org/10.1080/03004279.2015.963370

Sriraman, B. (2005). Are giftedness and creativity synonyms in mathematics? *Journal of Secondary Gifted Education, 17*(1), 20–36. https://doi.org/10.4219/jsge-2005-389

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ, 366,* l4898. https://doi.org/10.1136/bmj.l4898

Surmilasari, N., Marini, & Usman, H. (2022). Creative thinking with STEM-based project-based learning model in elementary mathematics learning. *Jurnal Pendidikan Dasar Nusantara, 7*(2), 434–444. https://doi.org/10.29407/jpdn.v7i2.17002

Suryaningsih, T., & Astuti, M. A. (2021). Pengaruh model pembelajaran open ended terhadap kemampuan berpikir kreatif matematis siswa kelas IV pada materi pecahan. *Elementar: Jurnal Pendidikan Dasar, 1*(1), 95–104. https://doi.org/10.15408/elementar.v1i1.20892

Tabach, M., & Friedlander, A. (2013). School mathematics and creativity at the elementary and middle-grade levels: How are they related? *ZDM Mathematics Education, 45*(2), 227–238. https://doi.org/10.1007/s11858-012-0471-5

Talkhan, E., Alhubaidah, S., Muthanna, A., & Qadhi, S. (2025). The effect of cooperative learning toward mathematics achievement of primary students: A systematic review using meta-analysis. *Social Sciences & Humanities Open, 12,* 102247. https://doi.org/10.1016/j.ssaho.2025.102247

Tlili, A., Saqer, K., Salha, S., & Huang, R. (2025). Investigating the effect of artificial intelligence in education (AIEd) on learning achievement: A meta-analysis and research synthesis. *Information Development, 41,* 825–842. https://doi.org/10.1177/02666669241304407

Turgut, S., & Turgut, I. G. (2018). The effects of cooperative learning on mathematics achievement in Turkey: A meta-analysis study. *International Journal of Instruction, 11*(3), 663–680. https://doi.org/10.12973/iji.2018.11345a

Ulfah, U., Prabawanto, S., & Jupri, A. (2017). Students' mathematical creative thinking through problem posing learning. *Journal of Physics: Conference Series, 895*(1), 012097. https://doi.org/10.1088/1742-6596/895/1/012097

Webb, N. M., & Mastergeorge, A. (2003). Promoting effective helping behavior in peer-directed groups. *International Journal of Educational Research, 39*(1–2), 73–97. https://doi.org/10.1016/S0883-0355(03)00074-0

Zhan, Z., He, L., & Zhong, X. (2024). How does problem-solving pedagogy affect creativity? A meta-analysis of empirical studies. *Frontiers in Psychology, 15,* 1287082. https://doi.org/10.3389/fpsyg.2024.1287082

TETY NUR CHOLIFAH is a doctoral researcher in the Faculty of Postgraduate, State University of Malang (Universitas Negeri Malang), Indonesia. Her research interests include elementary mathematics education, collaborative learning, and mathematical creative reasoning. Email: tety.nur.2321039@students.um.ac.id

BUDI EKO SOETJIPTO, PUNAJI SETYOSARI, and SRI UNTARI are faculty members in the Faculty of Postgraduate, State University of Malang (Universitas Negeri Malang), Indonesia, with research interests spanning instructional design, educational technology, and social studies and mathematics pedagogy.
