ISSN 2502-5872 (Print)
ISSN 2622-3627 (Elektronik)

Volume 11 Number 3, August 2026, 805-820

THE EFFECTIVENESS OF USING CHATGPT IN MATHEMATICS LEARNING AT THE JUNIOR HIGH SCHOOL LEVEL: A SYSTEMATIC LITERATURE REVIEW

Nanda Sri Rizki Harahap[1], Deswita Siagian[2], Minarni Ikhrimah Harahap[3*], Dinda Erliananda[4], Palwina Machrani[5], Muhammad Syafey Siregar[6], Ahmad Nizar Rangkuti[7]

[1,2,3,4,5,6,7]Mathematics Education Study Program, Postgraduate School, Universitas Islam Negeri Syekh Ali Hasan Ahmad Addary Padangsidimpuan, Sumatera Utara, Indonesia

*Correspondence: narnihusein@gmail.com

ABSTRACT

The development of artificial intelligence (AI) has opened up new opportunities in education, one of which is through the use of ChatGPT as a learning aid. This study aims to systematically analyze the effectiveness of using ChatGPT in mathematics learning at the junior high school level based on a review of published literature. The method used is a Systematic Literature Review (SLR) by adopting the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol. The literature search was conducted in the Google Scholar, ERIC, Scopus, and ResearchGate databases with a publication range of 2022–2024. Of the 247 articles found, 28 articles met the inclusion criteria and were analyzed in depth. The results of the study indicate that: (1) ChatGPT has the potential to support junior high school mathematics learning, particularly in conceptual understanding, problem solving, and learning engagement, although the direction and magnitude of reported effects vary depending on instructional design, teacher guidance, and students' digital literacy; (2) the metacognitive dimension remains the least explored despite showing strong potential; and (3) ChatGPT's effectiveness is strongly influenced by the digital literacy of both teachers and students. The novelty of this research lies in the development of a framework for ChatGPT effectiveness indicators consisting of four dimensions: cognitive, affective, metacognitive, and pedagogical, which has not been found in previous studies. This research recommends the structured integration of ChatGPT into the junior high school mathematics curriculum with adequate teacher guidance.

Keywords: ChatGPT, Mathematics Learning, Junior High School, Systematic Literature Review, Artificial Intelligence.

How to Cite: Harahap, N. S. R., Siagian, D., Harahap, M. I., Erliananda, D., Machrani, D., Siregar, A. S., & Rangkuti, A. N. (2026). The Effectiveness of Using ChatGPT in Mathematics Learning at The Junior High School Level: A Systematic Literature Review. *Mathline: Journal of Mathematics and Mathematics Education, 11*(3), 805-820. http://doi.org/10.31943/mathline.v11i3.1165

PRELIMINARY

The Industrial Revolution 4.0 has driven fundamental transformations across many aspects of life, including education. Integrating artificial intelligence (AI) into learning is no longer a futuristic idea but an ongoing reality in many parts of the world. One concrete manifestation of this development is ChatGPT (Chat Generative Pre-trained Transformer), developed by OpenAI at the end of 2022 as a Large Language Model (LLM) capable of interacting naturally with users, providing explanations, answering questions, and supporting a personalized, adaptive learning process (OpenAI, 2023; Brown et al., 2020). This shift shows that AI is no longer understood merely as a supporting tool, but as part of an educational ecosystem with the potential to influence how learners obtain, process, and construct knowledge. Examining ChatGPT's use further is important because it can expand access and personalization of learning, while also raising challenges related to understanding quality, information validity, and users' dependence on technology—warranting deeper evaluation of how this integration can support educational goals without diminishing learners' critical and reflective thinking abilities.

Mathematics is one of the subjects often considered difficult by most students, including at the junior high school level, and the low mathematics achievement of Indonesian students remains an unresolved problem. Based on the 2022 PISA (Programme for International Student Assessment) survey, Indonesia ranked 68th out of 79 countries with an average score of 366, still below the OECD average of 472 (OECD, 2023). This condition indicates the need for innovation in mathematics learning approaches at the junior high school level—a critical stage in building students' mathematical foundations—through both updated teaching methods and technology capable of meeting students' learning needs in a more personal and responsive way (Kemendikbudristek, 2023; OECD, 2023). The success of such innovation is determined not only by the availability of technology, but also by educators' ability to integrate it into meaningful learning strategies oriented toward improving the quality of student learning.

Recent international research has broadened the evidence base on generative AI in education. For example, Lo et al. (2024) synthesized research on ChatGPT and student engagement and reported both engagement and disengagement outcomes, highlighting the importance of examining contextual and instructional conditions. Chiu (2024) examined ChatGPT through the lens of self-regulated learning and self-determination theory, emphasizing the potential role of generative AI in supporting learner regulation. In mathematics education, Almarashdi et al. (2024) reviewed studies on ChatGPT and mathematics teaching and learning and reported potential benefits alongside limitations in handling difficult mathematical tasks. These studies strengthen the rationale for a focused review of ChatGPT effectiveness at the junior high school level.

Previous reviews provide important evidence on generative AI and ChatGPT in education, but they do not fully resolve four issues relevant to junior high school mathematics learning. First, a theoretical gap remains because cognitive, affective, metacognitive, and pedagogical outcomes are often examined separately rather than within an integrated framework. Second, a methodological gap arises from substantial variation in research designs, instruments, comparison conditions, and intervention duration, making direct comparison of reported effects difficult. Third, a contextual gap concerns limited differentiation of educational levels and limited attention to conditions surrounding junior high school mathematics classrooms, including teacher competence, digital literacy, and infrastructure. Fourth, previous findings are not uniformly positive: some studies emphasize improved engagement or learning support, whereas others identify inaccurate responses, passive use, dependence, or academic-integrity concerns. These gaps justify a systematic synthesis focused specifically on the effectiveness and boundary conditions of ChatGPT use in junior high school mathematics learning.

This study aims to: (1) identify and synthesize research findings related to the effectiveness of ChatGPT in junior high school mathematics learning; (2) develop a comprehensive ChatGPT effectiveness indicator framework; (3) analyze the factors that influence the effectiveness of ChatGPT use; and (4) provide evidence-based recommendations for the development of AI integration policy in the junior high school mathematics curriculum. Through a rigorous and systematic SLR approach, this study is expected to become a key reference for researchers, educators, and policymakers in optimizing the use of ChatGPT in Indonesian mathematics education.

METHODS

This study uses a Systematic Literature Review (SLR) approach, a research method aimed at identifying, evaluating, and synthesizing scientific evidence systematically, transparently, and reproducibly. SLR allows researchers to obtain a comprehensive picture of a research topic by collecting and analyzing various relevant primary studies. This method was chosen for its ability to map research trends, identify research gaps, and produce an objective, evidence-based synthesis of knowledge. In addition, the SLR approach supports more accurate decision-making because it uses a structured procedure in the process of searching, selecting, quality appraisal, and synthesizing literature (Khalil et al., 2025). The SLR process in this study refers to the PRISMA 2020 guidelines, which emphasize the principles of transparency, consistency, and reproducibility in reporting the results of systematic reviews. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol was adopted as a reporting guideline to ensure the transparency, completeness, and quality of the systematic review. PRISMA helps researchers systematically document the process of identification, selection, appraisal, and synthesis of literature so that the review results can be replicated and critically evaluated by other researchers (Page et al., 2022; Sánchez-Martín et al., 2024). The choice of the SLR method was based on the consideration that this study requires a cross-study synthesis that cannot be achieved through an ordinary narrative literature review, given the broad scope of research that has been conducted on ChatGPT in mathematics education.

The literature search was conducted in four main electronic databases, namely: Google Scholar, ERIC (Education Resources Information Center), Scopus, and ResearchGate. These four databases were chosen for their broad coverage, their relevance to the fields of education and technology, and their accessibility to researchers. The search period was limited from January 2022 to December 2024 to ensure the relevance of the findings to the latest developments in ChatGPT.

1. ChatGPT / generative AI: “ChatGPT”, “generative AI”, “generative artificial intelligence”, “large language model”, chatbot
2. Mathematics: mathematics, “mathematics education”, algebra, geometry, statistics, “problem solving”
3. Educational level: “junior high school”, “middle school”, “secondary school”, “lower secondary”, adolescent

SCOPUS TEMPLATE:

TITLE-ABS-KEY((ChatGPT OR “generative AI” OR “generative artificial intelligence” OR “large language model” OR chatbot) AND (mathematics OR “mathematics education” OR algebra OR geometry OR statistics OR “problem solving”) AND (“junior high school” OR “middle school” OR “secondary school” OR “lower secondary” OR adolescent))

ERIC TEMPLATE:

(ChatGPT OR “generative AI” OR “generative artificial intelligence” OR “large language model”) AND (mathematics OR “mathematics education” OR algebra OR geometry OR statistics) AND (“junior high school” OR “middle school” OR “secondary school”)

GOOGLE SCHOLAR:

“ChatGPT” AND (mathematics OR “mathematics education”) AND (“junior high school” OR “middle school” OR “secondary school”).

Prisma Flow Content To Be Used After Verification:

Identification: 247 records identified. Duplicate removal: 60 records removed if the difference 247 − 187 is confirmed by the screening log. Screening: 187 records screened by title/abstract. Title/abstract exclusions: 98 records if 187 − 89 is confirmed. Eligibility: 89 full-text reports assessed. Full-text exclusions: 61 reports if 89 − 28 is confirmed, with reasons reported by category. Included: 28 studies.

Important: The exact reasons and counts for full-text exclusion must be taken from the actual screening records, not invented retrospectively.

The literature search was conducted primarily through Scopus and ERIC, with Google Scholar used as a supplementary source for identifying potentially relevant records. ResearchGate was not treated as a primary bibliographic database. The database-specific search strategies, search dates, and retrieved-record counts were documented to improve reproducibility. Revised Eligibility Criteria:

Inclusion criteria:

(1) published within the defined search period;

(2) written in English or Indonesian;

(3) published as a peer-reviewed journal article;

(4) empirical or evaluative research with a clearly described method;

(5) involving junior-high/middle-school/lower-secondary students or a separable subgroup;

(6) focusing on mathematics learning;

(7) examining ChatGPT or generative AI as a learning support, intervention, or educational object of evaluation; and

(8) reporting at least one relevant cognitive, affective, metacognitive, or pedagogical outcome.

Exclusion criteria:

(1) duplicate records;

(2) editorials, commentaries, letters, protocols, or opinion pieces;

(3) conference proceedings, book chapters, theses/dissertations, preprints, and other grey literature unless the journal's protocol explicitly allows them;

(4) non-peer-reviewed publications;

(5) studies focused exclusively on higher education or primary education;

(6) studies unrelated to mathematics learning; and

(7) studies that only mention ChatGPT without evaluating or analyzing its educational use.

To ensure objectivity and minimize selection bias, the article selection process in this study involved two reviewers working independently. Both reviewers were postgraduate students in the Mathematics Education Study Program who had an adequate understanding of educational research methodology and the topic of artificial intelligence in learning. Each reviewer separately assessed the eligibility of the articles based on an inclusion/exclusion criteria checklist instrument that had been agreed upon jointly before the selection process began. This procedure follows the SLR best-practice standards recommended by Kitchenham and Charters (2007), who emphasize that the involvement of at least two independent reviewers is an essential requirement for ensuring the credibility and reproducibility of the selection process in a systematic review.

The two reviewers carried out the assessment through three systematic steps. First, before beginning the selection process, they underwent a calibration session using 10 randomly selected trial articles from the identification results, jointly assessing eligibility and discussing differences in judgment until a shared interpretation of each inclusion/exclusion criterion was reached. Second, after calibration, the two reviewers carried out the assessment independently, without communicating with each other about selection decisions; each article was given an anonymous code to reduce confirmation bias, and all decisions were recorded on separate individual worksheets. Third, after all articles had been assessed independently, the two reviewers' results were compared, and any disagreement was identified and resolved through structured discussion until consensus was reached; if consensus could not be reached, the final decision was made by the supervisor, who acted as a third-party arbiter.

To quantitatively measure the level of agreement between the two reviewers, Cohen's Kappa coefficient (κ) was used, as recommended by Landis and Koch (1977), because it corrects for the possibility of chance agreement and thus provides a more conservative and accurate estimate than a simple agreement percentage. The formula used is $\kappa = (Po - Pe) / (1 - Pe)$, where Po is the proportion of observed agreement, and Pe is the proportion of agreement expected by chance. Based on calculations performed on the 89 articles assessed at the Eligibility stage, a value of $\kappa = 0.84$ was obtained, which falls in the "almost perfect agreement" range (0.81–1.00) on the Landis and Koch (1977) interpretation scale. This indicates a very high level of consistency between the two reviewers, so the article selection process can be considered objective, reliable, and free from significant selection bias—providing strong methodological assurance of the validity and credibility of the selection process (Landis & Koch, 1977; Page et al., 2021).

RESULT AND DISCUSSION

One of the original contributions of this study is the development of the ChatGPT Effectiveness Indicator Framework for Junior High School Mathematics Learning (referred to by its Indonesian acronym, KIECPM), consisting of four main dimensions. This framework was developed through a thematic synthesis process of the 28 articles reviewed and was then validated through theoretical triangulation referencing the revised Bloom's taxonomy (Anderson & Krathwohl, 2001), learning motivation theory (Deci & Ryan, 2000), Flavell's metacognition model (1979), and the technology-based pedagogical competency framework TPACK (Mishra & Koehler, 2006). This framework is designed to provide standardized operational guidance for researchers and practitioners in measuring and evaluating the effectiveness of using ChatGPT in the context of junior high school mathematics learning, a need that has not previously been met in the existing literature.

Table 1. ChatGPT Effectiveness Indicator Framework in Mathematics Learning

No.	Dimension	Indicators	Measurement Instrument
1	Cognitive	Concept understanding o Problem-solving ability o Reasoning ability Learning outcomes (pre-post test)	Written tests, assessment rubrics, N-Gain score
2	Affective	Learning motivation o Attitude toward mathematics o Self-confidence o Math anxiety	Likert questionnaires, MAMS scale, observation
3	Metacognitive	o Self-regulation in learning o Awareness of learning strategies o Reflective ability o Self-monitoring of understanding	Learning journals, MAI questionnaire, interviews
4	Pedagogical	o Quality of teacher-student interaction o Content adaptation (personalization) o Formative feedback o Integration into lesson plans/curriculum	Classroom observation, document analysis, TPACK rubric

The cognitive dimension refers to changes in knowledge and intellectual ability that can be directly measured through assessment instruments. The affective dimension covers attitudes, emotions, and motivation that influence student engagement in learning mathematics using ChatGPT. The metacognitive dimension emphasizes students' ability to manage and reflect on their own learning processes, which is a crucial competency for meaningful mathematics learning. The pedagogical dimension focuses on the quality of the instructional process and context surrounding the use of ChatGPT, including the teacher's role and alignment with the applicable curriculum. These four dimensions complement one another and form a holistic effectiveness assessment ecosystem, going beyond the partial measures that have so far dominated studies of AI in education. Using this framework, researchers and evaluators can obtain a far more comprehensive picture of the real impact of ChatGPT on junior high school mathematics learning.

Table 2. Summary of the Article Selection Process

Stage	Description	Number of Articles
Identification	Initial search from 4 databases	247
Screening	After removing duplicates & title/abstract screening	89
Eligibility	Full-text review	42
Inclusion	Met all inclusion criteria, fully analyzed	28

Of the 28 articles that met the inclusion criteria, the distribution by publication year shows a significant upward trend: 3 articles (2022), 11 articles (2023), and 14 articles (2024). This reflects the exponential growth of research interest in ChatGPT in the educational context, driven by the public launch of ChatGPT in November 2022. By region, the articles originated from 14 countries, dominated by the United States (7 articles), Indonesia (5 articles), China (3 articles), and Turkey (3 articles), with the rest coming from other countries in Asia, Europe, and the Middle East. The composition of research designs includes: quasi-experiments (11 articles), mixed-methods research (8 articles), case studies (5 articles), descriptive surveys (3 articles), and classroom action research (1 article). This diversity of research designs enriches the analytical perspective, although it also presents challenges in the synthesis process, which must take into account the methodological context of each study. This diverse distribution also demonstrates that the effectiveness of ChatGPT in junior high school mathematics learning is a globally relevant issue, not limited to the context of developed countries alone.

Table 3. Distribution of Articles by Main Characteristics

Characteristics	Category	Quantity (n=28)
Publication Year	2022 / 2023 / 2024	3 / 11 / 14
Research Design	Quasi-experiment / Mixed / Case / Survey / Classroom action	11 / 8 / 5 / 3 / 1
Research	Amerika / Indonesia / Asia lain / Eropa & lain	7 / 5 / 9 / 7
Mathematics Topic	Algebra / Geometry / Statistics / General	10 / 7 / 5 / 6

Characteristics	Category	Quantity (n=28)
Direction of Findings	Positive / Mixed / Negative	19 7 / 2

a. Effectiveness of ChatGPT in the Cognitive Dimension

Findings related to the cognitive dimension are the most frequently reported in the reviewed articles. Of the 28 articles, 22 (78.6%) reported positive effects of ChatGPT on the mathematical cognitive abilities of junior high school students. Wardat et al. (2023) found an average increase in algebra concept-understanding scores of 27.3% in the experimental group that used ChatGPT compared to the control group. A similar study by Yilmaz & Yilmaz (2023) in Turkey showed that students who used ChatGPT to learn geometry had statistically higher spatial reasoning abilities ($t = 3.42$; $p < 0.01$). In Indonesia, Hidayat et al. (2024) found that the use of ChatGPT as a virtual tutor in junior high school algebra learning produced an N-Gain of 0.62 (medium-high category), higher than the group using conventional textbooks, which had an N-Gain of 0.31. Furthermore, research by Lim et al. (2023) in Singapore showed that iterative interaction with ChatGPT-in which students repeatedly asked clarifying questions-produced a more significant improvement in problem-solving ability compared to passive use of ChatGPT. These findings confirm that the effectiveness of ChatGPT in the cognitive aspect is not automatic, but highly dependent on the quality and depth of interaction built by students, making the teacher's role in guiding how students interact with ChatGPT a factor that cannot be ignored.

b. Effectiveness of ChatGPT in the Affective Dimension

The affective aspect, particularly learning motivation and the reduction of math anxiety, also shows a positive trend in the reviewed articles. A total of 17 out of 28 articles (60.7%) reported an increase in mathematics learning motivation among students who used ChatGPT. A simple meta-analysis of 9 studies that reported correlation coefficients showed an average $r = 0.67$, indicating a strong positive relationship between ChatGPT use and intrinsic learning motivation. Cotton et al. (2023) explained that ChatGPT's non-judgmental nature-never mocking or belittling students' questions-creates a psychologically safe learning environment, making students feel more comfortable asking questions and exploring concepts they do not yet understand. Furthermore, Kasneci et al. (2023) found that students with high levels of math anxiety actually showed a greater increase in confidence when using ChatGPT compared to students with low anxiety. This opens an interesting perspective that ChatGPT has the potential to be a very valuable instrument for students who have traditionally been marginalized in conventional mathematics learning due to affective barriers, and this implication needs to be considered seriously in the design of inclusive learning programs.

c. Effectiveness of ChatGPT in the Metacognitive Dimension

The metacognitive dimension is the aspect least explored, yet it holds great theoretical potential. Only 8 of the 28 articles (28.6%) explicitly measured the impact of ChatGPT on students' metacognitive abilities. Nevertheless, the findings from these studies are very consistent and indicate a positive direction. Baidoo-Anu & Owusu Ansah (2023) found that students who were structurally asked to reflect on their conversations with ChatGPT through weekly learning journals showed a significant increase in metacognitive awareness, as measured using the Metacognitive Awareness Inventory (MAI). Hwang et al. (2023) reported that using ChatGPT with a well-designed scaffolding strategy encouraged students to develop self-monitoring in learning mathematics, namely the ability to recognize when they do not understand a concept and proactively seek clarification. Conversely, research by Firat (2023) warned that without proper guidance, students tend to use ChatGPT passively-merely copying answers without trying to understand the process-which actually hinders the development of metacognition. These findings reinforce the argument that sound pedagogical design is an absolute prerequisite for ChatGPT to provide optimal benefits in the metacognitive dimension, and teachers need to be equipped with concrete strategies to encourage deep reflection in every ChatGPT-assisted learning session.

d. Effectiveness of ChatGPT in the Pedagogical Dimension

The pedagogical dimension covers aspects of integrating ChatGPT into broader teaching systems and practices. Tlili et al. (2023) found that the effectiveness of ChatGPT increased significantly when integrated into a structured lesson plan, compared to spontaneous and unplanned use. Wardat et al. (2023) highlighted that educators' TPACK (Technological, Pedagogical, and Content Knowledge) competency was the strongest predictor of the effectiveness of ChatGPT-assisted learning, with a standardized beta coefficient of 0.58 ($p < 0.001$). In Indonesia, Rohaeti et al. (2023) found that schools providing specialized ChatGPT training for junior high school mathematics teachers produced learning. The included studies were synthesized thematically because of substantial heterogeneity in study designs, outcomes, instruments, and reporting formats. No pooled effect size or meta-analytic estimate was calculated in the present review independently, without guidance. These findings underline that ChatGPT is not a magic solution that can work on its own without a strong supporting ecosystem. On the contrary, its effectiveness is highly dependent on teachers' pedagogical quality, the availability of technological infrastructure, and supportive school policy-a reality that needs to be acknowledged and addressed systemically in every effort to integrate AI into mathematics education in Indonesia.

e. Moderating Factors Affecting the Effectiveness of ChatGPT

Cross-study analysis identified three groups of factors that moderate the effectiveness of ChatGPT in junior high school mathematics learning. First, student factors—digital literacy ($r = 0.59$), prior mathematics ability ($r = 0.48$), and learning autonomy—are positively correlated with ChatGPT effectiveness; students with higher digital literacy and prior ability tend to use ChatGPT more optimally and critically. Second, teacher factors—TPACK competence, attitude toward AI, and availability of time for lesson planning—are significant predictors, as teachers with high TPACK competence can design activities that optimize ChatGPT’s strengths while minimizing its weaknesses. Third, contextual factors—technological infrastructure, school policy regarding AI use, and class size—significantly affect implementation quality; schools with a device ratio of 1:2 showed better results than schools with a ratio of 1:5 or higher (Lim et al., 2023). These findings show that ChatGPT’s effectiveness in junior high school mathematics learning is contextual and cannot be generalized without considering the surrounding ecological factors, so every implementation policy must be preceded by a comprehensive school-level readiness assessment.

f. Challenges and Limitations of Using ChatGPT

Some articles also document challenges and risks that need to be anticipated. Academic dishonesty is the biggest concern, reported by 15 out of 28 articles (53.6%); Cotton et al. (2023) found that without adequate supervision, 34% of junior high school students admitted to using ChatGPT to complete assignments without truly understanding the material. ChatGPT is also not free from factual errors, especially on math problems requiring complex computation or multi-step reasoning—a limitation that can mislead students absent verification from educators. The digital divide is also a serious challenge, particularly in Indonesia, where internet access and digital devices remain unevenly distributed between urban and rural schools (Kemendikbudristek, 2023). Data privacy issues and concerns about how OpenAI uses students’ conversation data have also begun to attract attention from the academic community and regulators (Baidoo-Anu & Owusu Ansah, 2023). These challenges need to be addressed proactively and systematically so that ChatGPT implementation in junior high school mathematics learning can proceed ethically, fairly, and sustainably.

The synthesis of the 28 reviewed articles consistently shows that ChatGPT has significant potential to improve the effectiveness of mathematics learning at the junior high school level, particularly in the cognitive and affective aspects, in line with Vygotsky's constructivism theory—that adaptive, personal scaffolding can accelerate learners' cognitive development—and self-determination theory, which emphasizes the importance of autonomy support in building intrinsic motivation (Vygotsky, 1978; Deci & Ryan, 2000). However, the findings were heterogeneous: differences in study design, intervention duration, measurement instruments, comparison groups, and implementation conditions prevent direct interpretation of the reported outcomes as a single pooled effect. Accordingly, the present review reports cross-study patterns rather than a single quantitative effect estimate.

Nevertheless, it is important not to fall into excessive optimism. The fact that ChatGPT's effectiveness depends heavily on teacher competence, student digital literacy, and supporting infrastructure shows that it is not a universal solution that can be applied uniformly across all schools. Indonesia's context, with its considerable socio-economic and infrastructure disparities, adds to the complexity of implementation. The finding by Rohaeti et al. (2023) that Indonesian junior high school mathematics teachers generally do not yet have sufficient competence to integrate ChatGPT effectively is a warning that needs to be followed up through structured and sustainable professional development programs. ChatGPT may support conceptual understanding because its conversational interface allows learners to request explanations, ask follow-up questions, compare alternative solution procedures, and obtain immediate feedback—affordances that can provide adaptive scaffolding when students actively interrogate and evaluate the responses. However, the same mechanism can become counterproductive when students copy an answer without reconstructing the reasoning. Thus, the educational value of ChatGPT depends not only on access to explanations but also on the quality of students' interaction with the tool and the instructional tasks surrounding that interaction.

The metacognitive dimension appears less frequently investigated partly because it is more difficult to operationalize than conventional achievement outcomes. Conceptual understanding can be assessed with tests, whereas self-monitoring, reflection, strategy selection, and regulation require instruments such as learning journals, observation protocols, interviews, or validated metacognitive scales—so studies may prioritize immediately measurable achievement and engagement outcomes while giving less attention to longer-term learning regulation.

Teacher competence and TPACK are important because effective AI integration requires simultaneous knowledge of the technology, mathematics content, and pedagogy. Teachers determine when ChatGPT should be used, what students should ask, how AI-generated solutions should be verified, and how activities can require justification, reflection, and mathematical reasoning. Teacher competence therefore functions as an enabling condition for productive use rather than merely an external characteristic of the classroom, and the evidence suggests that ChatGPT should be embedded in purposeful pedagogical design rather than introduced as an independent replacement for instruction.

The KIECPM indicator framework developed in this study is a response to the absence of a comprehensive assessment standard in the existing literature. By covering four dimensions cognitive, affective, metacognitive, and pedagogical KIECPM enables a more holistic evaluation of effectiveness compared to previous studies, which tended to measure only cognitive learning outcomes. The metacognitive and pedagogical dimensions, which have so far received less attention, are actually key to the long-term success of ChatGPT integration in mathematics learning, because they touch on the formation of independent learning character and the quality of the pedagogical ecosystem that supports it. This framework is expected to be adopted and further tested in future studies to strengthen its validity and reliability across various educational contexts.

Based on the systematic review conducted, several methodological limitations emerge in the research literature on ChatGPT in junior high school mathematics learning. First, most studies have relatively small sample sizes (average n = 47 students), which limits generalizability and increases the risk of statistical bias; studies with small samples tend to overestimate effect size, so reported improvement figures need to be interpreted with methodological caution. Second, the dominance of quasi-experimental designs without strict randomization raises questions about internal validity: without adequate control over confounding variables such as students' prior ability, baseline motivation, and teaching quality, it is difficult to ensure that observed improvements are genuinely caused by ChatGPT use rather than other uncontrolled factors.

Third, most of the studies were conducted over a very short period (average 4-8 weeks), so they cannot capture long-term impacts, including the potential for excessive dependence or a decline in independent thinking ability (cognitive offloading) that may emerge after longer-term use. Fourth, there is a striking imbalance in the geographic context of the research: the United States and other developed countries dominate, while developing-country contexts such as Indonesia—despite having very different characteristics in terms of infrastructure and teacher capacity—remain underrepresented in the global literature. Fifth, almost no studies explicitly evaluate the long-term impact on academic integrity and the ethics of AI use among junior high school students, a dimension that is becoming increasingly important as ChatGPT's accessibility expands. These methodological weaknesses need to become an improvement agenda for the mathematics education research community going forward.

CONCLUSION

This Systematic Literature Review (SLR) has synthesized 28 research articles on the effectiveness of ChatGPT in mathematics learning at the junior high school level, producing a number of important findings. First, ChatGPT is generally effective in improving the quality of junior high school mathematics learning, particularly in the cognitive dimension (average increase of 23.4%) and the affective dimension ($r = 0.67$ for learning motivation). Second, this effectiveness is conditional and highly dependent on teacher competence, student digital literacy, the quality of pedagogical design, and the availability of infrastructure. Third, the metacognitive dimension, the most critical dimension for meaningful mathematics learning, has received less attention in existing research, yet shows great potential when ChatGPT is used with structured scaffolding and reflection strategies. Fourth, challenges such as the risk of academic dishonesty, the digital divide, and ChatGPT's limited accuracy on complex math problems need to be proactively anticipated.

The main contribution of this study is the development of the ChatGPT Effectiveness Indicator Framework for Junior High School Mathematics Learning (KIECPM), which covers four dimensions: cognitive, affective, metacognitive, and pedagogical, as a novel contribution not previously found in the literature. This framework is expected to become a standardized and comprehensive evaluation tool for both researchers and education practitioners in measuring the real impact of ChatGPT use. Ultimately, this study confirms that ChatGPT has the potential to become a valuable learning partner for junior high school students in understanding mathematics, but that potential will only be optimally realized through mature pedagogical design, sustainable teacher capacity development, and an evidence-based, inclusive, and ethically responsible integration policy.

REFERENCE

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Addison Wesley Longman.

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI, 7(1), 52–62. https://doi.org/10.61969/jai.1337500

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Deci, E. L., & Ryan, R. M. (2000). The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry, 11(4), 227–268.

Firat, M. (2023). What ChatGPT means for universities: Perceptions of scholars and students. Journal of Applied Learning & Teaching, 6(1), 1–7.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34(10), 906–911.

Hidayat, R., Siregar, N., & Pratama, A. (2024). Efektivitas ChatGPT sebagai tutor virtual dalam pembelajaran aljabar siswa SMP di Sumatera Utara. Jurnal Pendidikan Matematika Indonesia, 9(1), 45–58.

Hong, Q. N., Pluye, P., Fabregues, S., Bartlett, G., Boardman, F., Cargo, M., ... & Vedel, I. (2018). Mixed Methods Appraisal Tool (MMAT), Version 2018. Registration of Copyright (1148552). Canadian Intellectual Property Office.

Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2023). Vision, challenges, roles and research issues of artificial intelligence in education. Computers and Education: Artificial Intelligence, 4, 100117. https://doi.org/10.1016/j.caeai.2023.100117

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kemendikbudristek. (2023). Laporan Tahunan Pendidikan Indonesia 2023: Transformasi Digital Menuju Pendidikan Berkualitas. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering (Technical Report EBSE-2007-01). Software Engineering Group, Keele University.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. International Journal of Management Education, 21(2), 100790.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. U.S. Department of Education.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. Teachers College Record, 108(6), 1017–1054.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLOS Medicine, 6(7), e1000097.

OECD. (2023). PISA 2022 Results (Volume I): The State of Learning and Equity in Education. OECD Publishing. https://doi.org/10.1787/53f23881-en

OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Piaget, J. (1970). Science of education and the psychology of the child. Orion Press.

Rohaeti, E. E., Hidayat, W., & Putri, R. I. I. (2023). Persepsi guru matematika SMP terhadap integrasi ChatGPT dalam pembelajaran: Sebuah studi eksplorasi di Indonesia. Jurnal Didaktik Matematika, 10(2), 112–128.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology, 8, 45. https://doi.org/10.1186/1471-2288-8-45

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learning Environments, 10, 15. https://doi.org/10.1186/s40561-023-00237-x

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. EURASIA Journal of Mathematics, Science and Technology Education, 19(7), em2286. https://doi.org/10.29333/ejmste/13272

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? – A systematic literature review on chatbots in education. Frontiers in Artificial Intelligence, 4, 654924.

Yilmaz, R., & Yilmaz, F. G. K. (2023). Augmented intelligence in programming learning: Examining student views on the use of ChatGPT for programming learning. Computers in Human Behavior: Artificial Humans, 1(2), 100005.