



Review Journal of Higher-Order Thinking and Metacognition

<https://rjhotm.litpam.com/index.php/rjhotm>



Higher-order thinking in STEM and science education: A systematic review of constructs, frameworks, and applications

Muhali ^{a,*}, Hunaepi ^b, Ni Nyoman Sri Putu Verawati ^c

^a Department of Science Education, Mandalika University of Education, Mataram, Indonesia

^b Doctoral Program in Educational Sciences, Universitas Pendidikan Ganesha, Singaraja, Indonesia

^c Department of Physics Education, University of Mataram, Mataram, Indonesia

ARTICLE INFO

Keywords:

Higher-order thinking
Critical thinking
Problem-solving
Metacognition
STEM education
Science education
Systematic review

ABSTRACT

This systematic review synthesizes how Higher-Order Thinking (HOT) is defined, theoretically framed, implemented, and assessed in STEM and science education. A Scopus search conducted on January 13, 2025 identified empirical qualitative, quantitative, and mixed-methods studies published between 2014 and 2024. Following PRISMA 2020 procedures, 24 studies met the predefined eligibility criteria and were included in the narrative and thematic synthesis. Methodological quality was appraised using the Mixed Methods Appraisal Tool, with appraisal findings used to calibrate the strength of interpretation rather than exclude eligible studies. The synthesis identified four recurring and overlapping HOT construct families: critical thinking and argumentation, problem-solving and scientific reasoning, creativity and innovation, and metacognition and self-regulation. Bloom's revised taxonomy was the most visible cognitive taxonomy, primarily guiding learning objectives and assessment design, while constructivist, inquiry-based, problem-based, discovery, and metacognitive models structured learning processes. No included study operationalized the SOLO Taxonomy as its primary empirical framework. Active and technology-supported approaches—including digital games, Internet of Things-supported investigation, educational robotics, flipped learning, online metacognitive instruction, and digital assessment—were generally associated with favorable HOT outcomes, although evidence strength varied because many studies used non-randomized, short-term, or context-specific designs. The review contributes an integrated analytical map linking HOT constructs, framework functions, learning activities, technologies, and assessment practices. It highlights the need for clearer construct-instrument alignment, stronger longitudinal and transfer designs, and sustained teacher professional development.

1. Introduction

1.1 Background and Rationale

Higher-Order Thinking (HOT) has become a central objective of contemporary education because learners increasingly need to interpret complex information, evaluate competing claims, generate defensible solutions, and apply knowledge in unfamiliar situations. These capabilities extend beyond factual recall and routine comprehension. HOT commonly involves analysis, evaluation, synthesis, creation, problem-solving, critical judgment, and the regulation of one's own thinking. However, HOT is not a single, uniformly defined ability. Its meaning varies across educational traditions, disciplines, theoretical frameworks, and assessment practices, creating persistent challenges for curriculum design, instructional implementation, and cross-study comparison.

* Corresponding author

E-mail address: muhali@undikma.ac.id (Muhali)

<https://doi.org/10.36312/rjhotm.v1i1.5>

Receive: June 2025; Revised: April 2026; Accepted: June 2026

ISSN 3164-0389/©2026 Muhali et al. Published by LITPAM. This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) License.

The literature frequently uses HOT or Higher-Order Thinking Skills (HOTS) as an umbrella term for several related but distinguishable constructs. Critical thinking concerns the interpretation and evaluation of information, the examination of assumptions, the assessment of evidence, and the construction of reasoned conclusions. Problem-solving involves representing a problem, generating and testing possible solutions, monitoring progress, and evaluating outcomes. Creativity emphasizes originality, flexibility, fluency, elaboration, and the production of alternative ideas. Metacognition refers to learners' awareness and regulation of their own cognition through planning, monitoring, strategy adjustment, and reflection. Argumentation, scientific reasoning, decision-making, and self-regulation also frequently appear as associated HOT processes. Although these constructs often co-occur in complex learning, treating them as interchangeable can obscure their distinct theoretical meanings and measurement requirements.

Previous reviews confirm this conceptual ambiguity. Darwis et al. (2024) identified substantial overlap between HOTS and critical thinking in problem-based learning research and emphasized the need for clearer operational definitions. Agustian et al. (2022) similarly recognized HOT as a major competence cluster in university chemistry laboratory education but reported variation in how it was defined and measured. Liu et al. (2021) mapped HOTS research into four broad facets—ability, instruction, curriculum, and learning—demonstrating the structural breadth of the field without establishing a single, universally accepted definition. Doolittle et al. (2023) also found that HOT-related outcomes were prominent in active-learning research, while the guiding constructs were often insufficiently defined. Collectively, these reviews show that conceptual fragmentation is not limited to isolated primary studies but is embedded in the broader HOT literature.

Some syntheses treat critical thinking, creativity, problem-solving, metacognition, and collaboration as discrete competencies within HOT or 21st-century learning frameworks (Parent et al., 2022; Tremblay & Poëllhuber, 2022; Smet et al., 2021). Other accounts emphasize their interdependence, presenting creativity as closely connected to problem-solving and metacognition and framing 21st-century competence as an integrated set rather than a collection of isolated skills (Parent et al., 2022; Tremblay & Poëllhuber, 2022). This tension between separation and integration has important consequences for educational research. Constructs must be distinguished sufficiently to support valid measurement, but their interaction must also be recognized because authentic learning tasks often require several HOT processes simultaneously.

1.2 Theoretical Foundations of Higher-Order Thinking

Bloom's Taxonomy remains one of the most influential frameworks for defining and organizing cognitive learning objectives. The original taxonomy was developed in 1956 and revised in 2001 to emphasize active cognitive processes. In its revised form, remembering, understanding, and applying are commonly distinguished from the more advanced processes of analyzing, evaluating, and creating. The taxonomy has substantially influenced curriculum design, instructional planning, and assessment development by providing educators with a structured vocabulary for specifying cognitive demands (Yanti & Thohir, 2024; Pujawan et al., 2022). Its continued prominence in HOT discourse reflects its practical value for aligning objectives, learning activities, and assessment items, particularly in science and STEM education (Rachmawati & Purwati, 2021; Katende, 2023).

Nevertheless, Bloom's Taxonomy should not be interpreted as a complete theory of how complex thinking develops. Its hierarchical representation may imply a fixed progression, although authentic reasoning is often recursive. Learners may recall information, generate a tentative interpretation, evaluate its adequacy, return to earlier evidence, and revise their conclusions repeatedly. Fixed cognitive hierarchies may therefore inadequately represent feedback loops, iterative reasoning, dialogic inquiry, and distributed cognition across learners, tools, and representations (Redshaw, 2024; Wilson et al., 2023). Bloom's Taxonomy is most useful when it is applied as a framework for specifying cognitive targets rather than as a rigid sequence through which all thinking must proceed.

The Structure of Observed Learning Outcomes (SOLO) Taxonomy offers a related but epistemologically different perspective. Whereas Bloom primarily classifies cognitive processes, SOLO characterizes the structural complexity of observable learning outcomes. Its five levels—pre-structural, uni-structural, multi-structural, relational, and extended abstract—describe movement from fragmented responses toward integrated understanding and transferable abstraction. This emphasis on the quality and organization of demonstrated knowledge makes SOLO useful for evaluating conceptual depth and progression (Kožuš et al., 2018; Triana et al., 2023; Sasson & Tifferet, 2025). Comparative work has identified partial correspondence between Bloom and SOLO, but the

frameworks differ in what they treat as evidence of higher-order learning: Bloom foregrounds cognitive operations, while SOLO foregrounds the integration and complexity of outcomes.

Both Bloom and SOLO are hierarchical frameworks, but they are limited when cognition is socially distributed or mediated through interaction with peers, language, digital tools, and cultural practices. Research on distributed creativity and collaborative cognition suggests that complex thinking may emerge across people, artifacts, dialogue, and shared representations rather than exclusively within an individual learner (Kimmel & van Alphen, 2025). Translanguaging, collaborative inquiry, and digitally mediated learning further demonstrate that knowledge construction is influenced by sociocultural and communicative conditions that cannot be fully captured by individual cognitive hierarchies (Itoi, 2024). These limitations do not invalidate Bloom or SOLO; instead, they indicate the need to combine cognitive-process and outcome frameworks with pedagogical, sociocultural, competency, and assessment perspectives.

The 21st Century Skills Framework broadens the scope of HOT by situating critical thinking, problem-solving, creativity, communication, collaboration, adaptability, and digital literacy within authentic social and professional contexts. Unlike taxonomies that primarily classify cognitive complexity, competency frameworks emphasize transferable performance across complex situations. They therefore respond to educational priorities extending beyond individual cognition to include interpersonal, technological, and contextual dimensions (Mutseekwa & Maeresera, 2024; Itoi, 2024; Jais et al., 2021; Ishmail & Ngene, 2024; Zana et al., 2024). Their breadth, however, creates operational difficulties because general competencies must be translated into discipline-specific tasks, observable indicators, and valid assessment criteria.

This epistemological contrast clarifies that cognitive taxonomies, competency frameworks, pedagogical models, instructional-design frameworks, and assessment frameworks should not be treated as equivalent. Cognitive taxonomies organize mental processes or the complexity of outcomes. Competency frameworks describe transferable capacities demonstrated in authentic and social contexts. Pedagogical frameworks guide how learning activities are organized, while instructional-design frameworks translate theoretical principles into sequences, supports, and learning resources. Assessment frameworks determine how performance is elicited, scored, and interpreted. Distinguishing these functions is essential for evaluating whether a study's theoretical claims, instructional design, and measurement strategy are coherently aligned.

1.3 Pedagogical Development of HOT

Problem-Based Learning (PBL) is among the most frequently used pedagogical approaches for fostering HOT. By presenting learners with complex or ill-structured problems, PBL requires them to identify what they know, determine what additional information is needed, evaluate possible explanations, and justify proposed solutions. Comparative studies have associated PBL with improvements in critical thinking, content knowledge, and problem-solving when compared with more conventional instructional approaches (Amin et al., 2020; Rehmat & Hartley, 2020). PBL can also strengthen learners' attitudes toward collaborative inquiry and lifelong learning by positioning knowledge as a resource for addressing meaningful problems rather than as content to be reproduced (Amin et al., 2020).

Inquiry-based learning supports HOT through related processes. It encourages learners to formulate questions, generate hypotheses, collect or examine evidence, interpret findings, and communicate conclusions. Technology-supported inquiry can additionally provide rapid feedback, simulations, and access to authentic data (Ruzaman & Rosli, 2020; Sotiriou et al., 2020). Project-based and hands-on learning extend these opportunities by requiring learners to design products, coordinate multiple sources of knowledge, and revise solutions in response to evidence or constraints (Chistyakov et al., 2023). These approaches are potentially valuable because they make reasoning visible through decisions, explanations, designs, and artifacts.

From a psychological perspective, metacognition provides a mechanism through which learners regulate these complex activities. Metacognitive learners plan their approach, monitor their understanding, identify difficulties, evaluate strategy effectiveness, and revise their reasoning. Integrating metacognitive prompts into instruction can therefore strengthen critical thinking and problem-solving by encouraging learners to examine how they reached a conclusion rather than focusing only on whether the answer was correct (Maryani et al., 2022; Marthaliakirana et al., 2022). Metacognition also supports independence because learners become more capable of selecting and adapting strategies without continuous teacher direction.

Computational thinking introduces another perspective on HOT by emphasizing decomposition, abstraction, pattern recognition, algorithmic reasoning, modeling, and iterative problem-solving. These processes can support analytical reasoning in science and STEM when they are embedded in meaningful disciplinary problems rather than taught only as technical procedures (Gouvea, 2023). Nevertheless, computational thinking should not automatically be equated with HOT. Its higher-order character depends on whether learners merely execute predetermined steps or independently model problems, evaluate alternatives, and revise solutions.

1.4 Challenges in Implementation and Assessment

Despite widespread recognition of HOT, implementation remains inconsistent. One source of difficulty is definitional ambiguity. Some educators prioritize critical thinking and problem-solving, while others foreground creativity, innovation, communication, or metacognition. These differences influence learning objectives, activity design, and assessment selection (Kanphukiew & Nuangchalerm, 2024). Without explicit construct definitions, studies may use the same HOTS label while measuring substantially different outcomes, or they may use different labels for comparable reasoning processes.

Assessment is a particularly important challenge. Standardized and classroom assessments often reward factual recall, procedural reproduction, or single correct answers, which can discourage teachers from allocating instructional time to inquiry, argumentation, creativity, and complex problem-solving. Traditional tests may also underrepresent the recursive and contextual nature of HOT (Sotiriou et al., 2020). A test described as measuring HOTS may assess only critical thinking, problem-solving, achievement, or one subset of Bloom's higher cognitive levels. Valid HOT assessment therefore requires clear construct definitions, tasks that elicit the intended processes, and scoring procedures aligned with those processes.

Teacher preparedness further shapes implementation. Educators need sufficient knowledge to define HOT constructs, design cognitively demanding questions, facilitate evidence-based discussion, and assess complex performance. However, professional development frequently focuses on adopting a teaching method or digital tool without addressing construct definition, instructional alignment, and assessment validity. Kinoshita (2022) showed that teachers' interpretations of critical thinking affected how it was enacted in science classrooms. Chung (2019) also illustrated that active learning can support critical thinking and problem-solving, but effective implementation depends on structured facilitation, contextualized tasks, and opportunities for self-directed learning.

Cultural and institutional contexts also matter. Curriculum mandates, assessment regimes, class size, teacher autonomy, resource access, and expectations about classroom participation can either facilitate or constrain HOT-oriented instruction. Systems emphasizing high-stakes factual examinations may create tensions between curriculum aspirations and classroom practice. However, broad cultural generalizations should be avoided because implementation is shaped by interactions among policy, institutional capacity, teacher preparation, and local pedagogical traditions rather than by culture alone.

1.5 Technology and HOT Development

Digital technologies have expanded the range of environments in which HOT can be developed. Simulations, digital games, mobile inquiry tools, collaborative media, online platforms, robotics, and interactive assessment systems can support experimentation, feedback, evidence interpretation, decision-making, and iterative problem-solving (Afikah et al., 2022; Reyna & Meier, 2020). Their educational value, however, depends on the cognitive and metacognitive activities they require. Technology does not promote HOT merely because it is interactive or engaging.

Augmented reality and virtual reality have been proposed as tools for creating immersive and experiential learning environments (Wrzesien & Alcañiz Raya, 2010; Zhang & Wang, 2021). Nevertheless, immersive presence or improved engagement should not be treated as direct evidence of critical thinking, creativity, problem-solving, or metacognition. Technology-enhanced HOT claims require explicit alignment among the technology's features, the intended cognitive process, the learning task, and the assessment instrument. For example, a simulation may support hypothesis testing, a digital game may support strategic decision-making, and a collaborative platform may support argumentation, but these outcomes must be measured directly.

The same principle applies to inquiry-based digital learning and co-creation through mobile technologies. Digital environments can extend opportunities for learners to construct and communicate knowledge, but their contribution depends on the quality of the task, feedback, scaffolding, and learner agency (Reyna & Meier, 2020). Consequently, the review of technology-enhanced HOT should examine not only which tools were used but also how their design features were theoretically connected to cognitive, metacognitive, or collaborative outcomes.

1.6 Research Gap and Review Rationale

The fragmented conceptualization of HOT presents a cumulative problem for educational research. Existing studies and reviews often examine one component of the field, such as critical thinking within PBL, HOT assessment, active learning, chemistry laboratory competence, a specific taxonomy, or a particular technology (Agustian et al., 2022; Darwis et al., 2024; Doolittle et al., 2023; Liu et al., 2021). These contributions are valuable, but they do not fully reconcile how HOT is defined, which type of framework is being used, how learning tasks operationalize the construct, and whether assessment instruments measure the claimed outcome.

Previous work also varies in whether HOT-related constructs are treated separately or as an integrated competence set. Some reviews catalogue critical thinking, creativity, problem-solving, collaboration, and metacognition as distinct dimensions (Parent et al., 2022; Tremblay & Poëllhuber, 2022; Smet et al., 2021), while others emphasize their overlap and interdependence. This variation indicates the need for a synthesis that preserves construct boundaries while explaining how the constructs interact in authentic educational activities.

Theoretical ambiguity further complicates comparison. Bloom's Taxonomy, SOLO, 21st-century competency frameworks, constructivism, PBL, inquiry learning, instructional-design models, and assessment rubrics are often discussed together as "frameworks," even though they operate at different analytical levels. Without distinguishing their functions, it becomes difficult to determine whether a study uses a framework to define cognition, describe learning progression, organize instruction, justify a technology, or evaluate performance.

Methodological limitations also remain important. Many HOT interventions use short-term pretest–posttest designs, small or context-specific samples, and heterogeneous instruments. Consequently, reported improvements may reflect immediate task performance rather than sustained development or transfer to new contexts (Doolittle et al., 2023; Liu et al., 2021). Study-by-study descriptions alone cannot resolve these issues. A structured synthesis is needed to compare construct emphasis, framework function, intervention design, assessment alignment, and methodological quality across studies (Chang et al., 2020).

1.7 Novelty, Contributions, and Objectives

The novelty of this review lies in integrating three levels of analysis that are often examined separately: the constructs used to define HOT, the theoretical or practical frameworks used to interpret it, and the pedagogical or technological applications through which it is implemented. Rather than treating HOT as a single undifferentiated outcome, the review maps study-specific terminology onto recurring construct families while retaining distinctions among critical thinking and argumentation, problem-solving and scientific reasoning, creativity and innovation, and metacognition and self-regulation. This approach addresses the conceptual overlap highlighted in previous research (Darwis et al., 2024; Liu et al., 2021).

The review also distinguishes framework types according to their functions. Bloom's Taxonomy is examined as a cognitive-target and assessment-planning framework; SOLO as a theoretical framework for the depth and integration of learning outcomes; 21st-century frameworks as competency-oriented accounts of transferable performance; constructivist, inquiry, problem-based, and project-based approaches as pedagogical models; and rubrics or tests as assessment structures. This functional classification clarifies boundaries that are frequently blurred in HOT research and supports more precise comparisons among studies.

A further contribution is the use of explicit alignment criteria to examine technology-supported learning. Alignment is evaluated through the relationship among the stated learning objectives, the cognitive or metacognitive demands of the task, and the assessment instrument used to measure the outcome. This prevents technological novelty, immersion, or engagement from being interpreted automatically as evidence of HOT development (Afikah et al., 2022; Reyna & Meier, 2020; Wrzesien & Alcañiz Raya, 2010).

The review extends previous work by aggregating cross-study patterns instead of relying solely on narrative summaries of individual interventions. It examines recurring construct combinations, framework functions, educational levels, pedagogical approaches, technology categories, assessment practices, and methodological limitations. This comparative structure provides educators with guidance for aligning learning objectives, activities, and assessment; assists researchers in improving construct comparability; and offers technology developers criteria for connecting digital features with intended cognitive processes (Müller & Pietzner, 2020).

Accordingly, this systematic review aims to identify how HOT is defined and operationalized in empirical STEM and science education research; examine the theoretical, competency, pedagogical, instructional-design, and assessment frameworks used to support HOT; analyze conceptual overlaps and distinctions among critical thinking, argumentation, problem-solving, scientific reasoning, creativity, innovation, metacognition, and self-regulation; compare how framework functions are translated into learning tasks and assessment practices; and formulate recommendations for theoretically aligned, methodologically rigorous, and context-responsive HOT research and educational implementation.

2. Method

2.1 Protocol and Reporting Standard

No review protocol was prospectively registered or deposited in a public repository. The review was reported using the PRISMA 2020 framework, particularly its identification, screening, retrieval, eligibility-assessment, and inclusion stages. Although the review was not preregistered, the search strategy, eligibility criteria, screening decisions, full-text exclusion reasons, data-extraction fields, quality-appraisal procedure, and synthesis approach were documented to improve transparency and reproducibility. The absence of prospective protocol registration is acknowledged as a methodological limitation because the completed review procedures cannot be compared with a publicly available prespecified protocol. Transparent documentation of screening rules, reviewer decisions, inter-reviewer agreement, and deviations from planned procedures is recommended to reduce selection and analytical bias in systematic reviews (Fleming et al., 2023; Mayo-Wilson et al., 2021; López-Nicolás et al., 2021; Gard et al., 2025; Mak et al., 2025).

2.2 Study Design

This systematic literature review synthesized empirical research on Higher-Order Thinking (HOT) in STEM and science education. The review examined how HOT was defined, operationalized, theoretically framed, implemented through pedagogical and technology-supported approaches, and assessed across educational contexts. The review was limited to studies published between 2014 and 2024 to capture contemporary developments in the field. Eligible publications comprised qualitative, quantitative, and mixed-methods empirical studies published in peer-reviewed journals or as full conference proceeding papers.

The scope was restricted to STEM and science education because these fields frequently require learners to analyze evidence, construct explanations, solve complex problems, evaluate competing claims, make decisions, and regulate their thinking. The review therefore focused not only on whether a study used the terms HOT or HOTS, but also on the specific constructs, learning processes, frameworks, interventions, and assessment indicators through which higher-order thinking was represented.

2.3 Search Strategy

The literature search was conducted exclusively in the Scopus database on January 13, 2025. Scopus was selected because of its multidisciplinary indexing and standardized bibliographic metadata across education, psychology, science, technology, engineering, and mathematics. Its use supported a consistent search and screening process across the interdisciplinary domains included in the review. Nevertheless, reliance on a single database may have reduced the comprehensiveness of the evidence base by excluding studies indexed only in sources such as Web of Science or ERIC. This limitation was considered when interpreting the scope and transferability of the findings.

The search strategy was designed to capture three dimensions: the HOT construct family, the presence of conceptual or theoretical structuring, and the educational domain. The following Boolean query was applied to titles, abstracts, and keywords:

TITLE-ABS-KEY (((("Higher-Order Thinking" OR "Critical Thinking" OR "Problem-Solving") AND ("Framework" OR "Taxonomy" OR "Education" OR "Learning"))) AND ("Science Education" OR "STEM Education" OR "Learning"))

The query combined terms representing HOT, critical thinking, and problem-solving with terms related to frameworks, taxonomies, education, and learning. It also required an explicit connection to science education, STEM education, or learning. This structure was intended to balance sensitivity by retrieving studies addressing major HOT-related constructs and specificity by requiring relevance to conceptual frameworks or educational implementation.

The initial search identified 1,187 records. Scopus database filters were then applied to retain publications issued between 2014 and 2024, publications written in English, and documents classified as either articles or conference papers. Application of these filters removed 356 records before screening and left 831 records for title and abstract screening. The query did not include all conceptually adjacent HOT terms, such as creativity, metacognition, self-regulation, argumentation, scientific reasoning, analytical thinking, or decision-making, as independent search terms. These constructs were coded when they appeared in studies retrieved through the specified search strategy. The possibility that studies using only these adjacent labels were not retrieved is acknowledged as a search-term limitation.

2.4 Eligibility Criteria

The eligibility criteria were established to determine whether each publication was substantively and methodologically relevant to the review objectives (see Table 1). These criteria were used to establish eligibility and were not interpreted as evidence that an included study was necessarily of high methodological quality. Methodological quality was evaluated separately using the Mixed Methods Appraisal Tool.

Table 1. Inclusion and exclusion criteria

Eligibility criterion	Inclusion	Exclusion
Focus	Studies examining Higher-Order Thinking, higher-order thinking skills, critical thinking, problem-solving, or related constructs, with an explicit connection to definitions, operational indicators, frameworks, pedagogical applications, technology-supported learning, or assessment.	Studies unrelated to HOT or studies mentioning a HOT-related construct only incidentally without examining its definition, operationalization, framework, application, or assessment.
Study type	Qualitative, quantitative, or mixed-methods empirical research published in a peer-reviewed journal or as a full conference proceeding paper.	Systematic reviews, literature reviews, bibliometric studies, conceptual or theoretical papers without empirical data, editorials, opinion papers, book chapters, conference abstracts, posters, and conference papers without sufficient methodological or empirical information.
Language	Publications whose full texts were written in English.	Publications whose full texts were written in languages other than English.
Publication range	Publications issued between 2014 and 2024.	Publications issued before 2014 or after 2024.
Educational context	Studies conducted in STEM or science education across primary, secondary, tertiary, teacher-education, or closely related educational settings.	Studies conducted exclusively in non-STEM contexts or outside an educational setting.
Empirical relevance	Studies reporting extractable empirical evidence on HOT definitions, operational indicators, learning processes, instructional approaches, technology-supported applications, assessment practices, or learning outcomes.	Studies that did not provide extractable empirical evidence relevant to the review objectives.

Eligibility criterion	Inclusion	Exclusion
Reporting completeness	Full journal articles or full conference papers reporting sufficient information about participants or data sources, procedures, analyses, and findings.	Abstract-only records, posters, presentation summaries, or reports lacking sufficient methodological or empirical information for analysis.

Full-text availability was treated as a report-retrieval issue rather than as a substantive eligibility criterion. Reports that could not be obtained after the documented retrieval attempts were classified as reports not retrieved and were not included in the full-text eligibility assessment.

2.5 Study Selection Process

The study-selection process followed the PRISMA stages of identification (Figure 1), title and abstract screening, report retrieval, full-text eligibility assessment, and final inclusion. The Scopus search identified 1,187 records. No duplicate records were identified. Application of the publication-year, English-language, and document-type filters removed 356 records before screening, leaving 831 records for title and abstract screening.

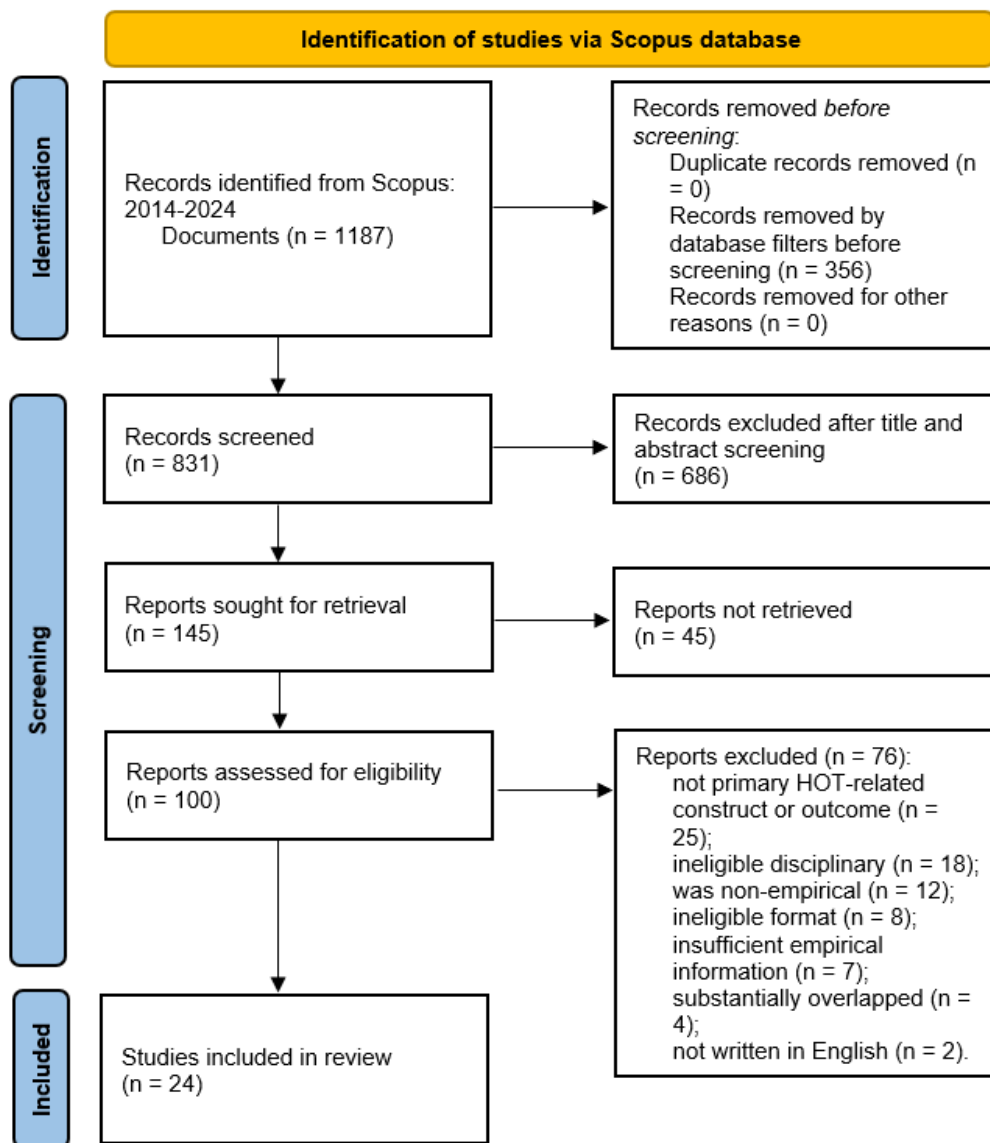


Figure 1. Identification of studies through Scopus

During title and abstract screening, 686 records were excluded because their topics, publication characteristics, study designs, or educational contexts did not meet the eligibility criteria. Full texts were sought for the remaining 145 reports. Searches were conducted through publisher websites, institutional library access, DOI-based searches, institutional repositories, and other publicly available scholarly sources. Bibliographic records or article pages for several unavailable reports were also identified on ResearchGate, but downloadable full-text PDF files were not provided. Full-text requests were submitted through ResearchGate; however, no responses were received. Consequently, 45 reports were classified as not retrieved and were neither assessed for full-text eligibility nor included in the synthesis.

The remaining 100 reports underwent full-text eligibility assessment. Seventy-six reports were excluded using one primary exclusion reason for each report to avoid double counting. The excluded reports comprised 25 publications in which the HOT-related construct was not examined as a primary construct or outcome, 18 publications from ineligible educational or disciplinary contexts, 12 non-empirical publications, eight publications with ineligible or incomplete formats, seven reports with insufficient empirical information for data extraction, four duplicate or substantially overlapping reports, and two reports whose full texts were not written in English. Following this process, 24 empirical studies met all eligibility criteria and were included in the final synthesis.

The final study-selection figures were therefore 1,187 records identified, 356 records removed by database filters, 831 records screened, 686 records excluded after title and abstract screening, 145 reports sought for retrieval, 45 reports not retrieved, 100 reports assessed for full-text eligibility, 76 full-text reports excluded, and 24 studies included in the review.

2.6 Reviewer Involvement and Inter-Reviewer Agreement

Two reviewers were involved in the assessment of the 145 reports that remained after title and abstract screening. They verified retrieval status and independently evaluated the 100 reports for which complete full texts were available. Both reviewers applied the same eligibility criteria and used a structured screening checklist to record their initial inclusion or exclusion decisions before consensus discussion.

The reviewers initially agreed on 92 of the 100 full-text eligibility decisions, corresponding to an overall agreement of 92.0%. Cohen's kappa was 0.80, indicating substantial agreement beyond chance. The reviewers disagreed on eight reports. In three cases, Reviewer 1 recommended inclusion while Reviewer 2 recommended exclusion; in five cases, Reviewer 2 recommended inclusion while Reviewer 1 recommended exclusion. All disagreements were resolved through discussion and reference to the predefined eligibility criteria, and no third adjudicator was required. The final consensus resulted in 24 included and 76 excluded reports.

No formal pilot screening exercise was conducted before the assessment of the 145 reports. This is acknowledged as a procedural limitation because pilot screening can reveal ambiguities in eligibility criteria before full assessment. Nevertheless, the two reviewers applied identical decision rules, documented their initial judgments, and recorded a single primary exclusion reason for each excluded full-text report. Independent assessment, explicit disagreement-resolution procedures, predefined decision rules, and quantitative reporting of inter-reviewer agreement are recommended to strengthen the transparency of systematic-review screening (Gard et al., 2025; Fleming et al., 2023; Cumming et al., 2023; Sandoval-Lentisco et al., 2025; López-Nicolás et al., 2021; Mak et al., 2025).

2.7 Data Extraction

Data were extracted using a standardized matrix designed to ensure consistency across the 24 included studies. The matrix recorded bibliographic information, publication year, publication type, country or geographical setting, educational level, disciplinary context, participant characteristics, sample size, research design, data-collection procedures, analytical methods, and intervention duration where applicable. It also captured each study's definition or description of HOT, the specific HOT constructs examined, the conceptual or theoretical structures referenced, the pedagogical intervention, the use of technology, the assessment instruments, the principal findings, the direction of the reported outcomes, and the limitations identified by the primary-study authors.

The construct fields included critical thinking, argumentation, problem-solving, scientific reasoning, creativity, innovation, metacognition, self-regulation, analytical thinking, and decision-making. The framework

fields distinguished cognitive taxonomies, learning or epistemological theories, competency frameworks, pedagogical models, instructional-design models, and assessment frameworks. This distinction prevented conceptually different structures, such as Bloom's Taxonomy, constructivism, problem-based learning, ADDIE, and assessment rubrics, from being treated as interchangeable forms of framework.

The extraction matrix also recorded whether each study explicitly referenced Bloom's Taxonomy, the SOLO Taxonomy, or a 21st Century Skills Framework. Framework alignment was coded as explicit when the primary-study authors directly named and applied the framework. Alignment was coded as inferred when the framework was not named but the learning objectives, task structure, or assessment indicators corresponded to its defining characteristics. Inferred alignment was interpreted cautiously because it represented an analytical judgment made during the review rather than a theoretical claim made by the primary-study authors.

2.8 Methodological Quality Appraisal

The methodological quality of the 24 included studies was assessed using the Mixed Methods Appraisal Tool (MMAT). MMAT was selected because the corpus included qualitative studies, quantitative randomized and non-randomized studies, quantitative descriptive studies, and mixed-methods research. Its design-specific criteria enabled the studies to be evaluated according to the methodological requirements appropriate to their respective research designs rather than by applying one undifferentiated standard across the entire evidence base (Liu & Mehigan, 2021).

One reviewer conducted the initial research-design classification and MMAT appraisal of all 24 included studies, while a second reviewer independently verified each judgment against the full-text reports. Disagreements concerning design classification or individual MMAT criteria were resolved through discussion and reference to the MMAT guidance until consensus was reached. Each design-specific criterion was recorded as "Yes," "No," or "Cannot tell." No overall numerical quality score was calculated; instead, the criterion-level judgments were used to determine the appropriate interpretive weight of each study in the synthesis.

The appraisal examined the clarity and appropriateness of the research questions, the suitability of the design, the adequacy of sampling and recruitment, the validity and reliability of the instruments, the comparability of intervention groups where relevant, the handling of confounding variables, the completeness of outcome data, the appropriateness of statistical analyses, the transparency of qualitative data collection and interpretation, and the integration of qualitative and quantitative components in mixed-methods studies. Because the review focused specifically on HOT, the appraisal also considered whether the instrument or performance task was adequately aligned with the HOT construct claimed by the primary-study authors.

Methodological quality was not used as an additional eligibility criterion. All 24 studies that met the predefined inclusion criteria were retained in the synthesis regardless of their appraisal findings. MMAT judgments were recorded at the study level and used to qualify the strength, limitations, and appropriate interpretive weight of the evidence. Studies with methodological limitations contributed to the descriptive and thematic synthesis, but they were not used as the sole basis for strong causal or effectiveness claims. Consistent with MMAT guidance, the appraisal results were not combined into a single numerical quality score.

2.9 Data Synthesis

The studies differed substantially in research design, participant characteristics, educational level, disciplinary context, intervention type, duration, technology use, assessment instruments, and outcome reporting. Because of this heterogeneity, statistical pooling through meta-analysis was not considered appropriate. The evidence was therefore integrated using narrative synthesis complemented by thematic synthesis. Narrative synthesis was used to compare findings across qualitative, quantitative, developmental, and mixed-methods studies, while thematic synthesis was used to identify recurring definitions, constructs, frameworks, learning mechanisms, assessment practices, and contextual patterns (McMahon et al., 2022).

The synthesis followed a hybrid inductive–deductive approach. The inductive component identified recurring construct labels, descriptions, learning processes, and assessment indicators in the included studies. The deductive component organized the evidence according to predefined categories of HOT construct, framework type,

pedagogical approach, technology use, educational context, assessment method, research design, and outcome direction.

A two-layer coding procedure was used to address definitional ambiguity. In the first layer, each study's definition or description of HOT or HOTS was coded into one or more recurring construct families. These families comprised critical thinking and argumentation, problem-solving and scientific reasoning, creativity and innovation, and metacognition and self-regulation. Multiple coding was permitted when a study explicitly examined more than one HOT construct. The synthesis distinguished construct emphasis from construct coverage. Construct emphasis referred to the HOT construct explicitly identified by the primary-study authors, whereas construct coverage referred to additional HOT processes represented in the learning tasks, intervention mechanisms, or assessment indicators.

In the second layer, each study was coded according to the type and function of the theoretical, conceptual, pedagogical, instructional, or assessment structure it used. Cognitive taxonomies, learning and epistemological theories, competency frameworks, pedagogical models, instructional-design models, and assessment frameworks were analyzed as distinct categories. Framework alignment was evaluated through the relationship among the stated learning objectives, the structure of the learning tasks or interventions, and the assessment instruments or performance indicators used to measure HOT outcomes.

Frequencies and percentages were calculated where appropriate to describe the distribution of constructs, framework types, research designs, educational levels, geographical contexts, pedagogical approaches, technology categories, and assessment methods. Because studies could receive more than one construct or framework code, multiple-response percentages were allowed to exceed 100%. The coded data were then compared across educational levels, disciplines, countries, research designs, intervention types, and methodological-quality profiles to identify recurring patterns, co-occurrences, divergences, and gaps in the evidence base.

All 24 included studies contributed to the synthesis. However, their evidentiary roles were differentiated according to research design and methodological quality. Comparative intervention studies were used to examine reported differences in HOT outcomes, while one-group, descriptive, correlational, qualitative, developmental, and needs-analysis studies were used to explain conceptualizations, implementation processes, design requirements, assessment approaches, and contextual factors. Studies reporting only product validation or feasibility were not treated as evidence of intervention effectiveness unless they also reported appropriate empirical outcome data.

Only the 24 included empirical studies contributed evidence to the Results section. Prior systematic reviews, theoretical publications, methodological sources, policy documents, and other contextual literature were used separately in the Introduction or Discussion and were not counted as findings derived from the included empirical corpus.

3. Results

3.1 Overview of the Included Studies

The review included 24 empirical studies examining Higher-Order Thinking (HOT) in STEM and science education. Table 2 summarizes the study contexts, research designs, primary HOT constructs, pedagogical or technological approaches, and contributions to the synthesis. Appendix A provides the complete extraction matrix, including participant characteristics, instruments, analytical procedures, study limitations, coding decisions, and study-level Mixed Methods Appraisal Tool (MMAT) judgments.

The studies covered primary, secondary, tertiary, and teacher-education settings and employed quantitative non-randomized, quantitative descriptive, qualitative, developmental, and mixed-methods designs. The corpus included comparative interventions, one-group pretest–posttest studies, descriptive and correlational investigations, qualitative studies, product-development research, feasibility studies, and needs analyses.

All 24 studies were retained because they met the predefined eligibility criteria. MMAT appraisal was used to assess the methodological strengths and limitations of each study rather than as an additional exclusion criterion. Comparative intervention studies contributed evidence on reported differences in HOT outcomes, while descriptive, qualitative, developmental, and needs-analysis studies contributed evidence on conceptualization, implementation, assessment, and contextual conditions.

Table 2. Characteristics and synthesis of the 24 studies included in the review

Study	Context and design	Primary HOT focus	Pedagogical, theoretical, or technological approach	Main contribution to the synthesis
Kinoshita (2022)	Elementary science teachers in Japan; descriptive study	Critical thinking	Science-teaching practices and teacher interpretation of critical thinking	Provides contextual evidence on how teachers understand and attempt to incorporate critical thinking into elementary science instruction.
Bolat and Karamustafaoğlu (2024)	Science teachers; professional-development study	HOTS assessment	Training in constructing HOTS-oriented science questions	Reports improvement in teachers' ability to design questions that require analysis, evaluation, and higher-level reasoning.
Amin et al. (2020)	Secondary-level science students; quasi-experimental study	Critical thinking and environmental attitudes	Problem-based learning	Indicates that problem-based learning was associated with stronger critical-thinking outcomes than conventional instruction.
Suwardika et al. (2024)	Higher-education students; intervention study	Critical thinking, self-regulation, and communication	Flipped classroom supported by whiteboard animation and learning modules	Shows how a technology-supported flipped model can integrate cognitive, metacognitive, and communication outcomes.
Fathonah et al. (2024)	227 higher-education students; comparative non-randomized study	Critical and creative thinking	Multirepresentation-Based Creative Problem-Solving model	Reports greater gains in critical and creative thinking when problem-solving activities incorporated verbal, visual, and mathematical representations.
Maryani et al. (2022)	119 preservice elementary teachers; development and comparative experimental study	HOTS and metacognition	Metacognition-integrated online science learning	Demonstrates that planning, monitoring, evaluation, and reflection can be embedded in synchronous and asynchronous science learning to support HOTS.
Sidiq et al. (2021)	60 fifth-grade students; quasi-experimental study	Critical thinking	Habituation with HOTS-based science questions	Reports higher post-intervention critical-thinking performance among students regularly exposed to HOTS-oriented science questions.
Darmaji et al. (2024)	180 secondary students; comparative quantitative study	Critical thinking and science-process skills	Generic science skills and gender comparison	Provides descriptive evidence on relationships among generic skills, science-process skills, critical thinking, and gender, although reporting inconsistencies limit strong interpretation.
Kanphukiew and Nuangchalem (2024)	30 eighth-grade students; pre-experimental study	Scientific problem-solving	Active science learning grounded in constructivism	Reports post-intervention achievement and problem-solving scores above predetermined criteria, but the absence of a comparison group limits causal interpretation.

Study	Context and design	Primary HOT focus	Pedagogical, theoretical, or technological approach	Main contribution to the synthesis
Zulfiani et al. (2023)	32 secondary students, one teacher, and 10 experts; research and development	HOTS assessment and scientific reasoning	MathSci 21st Android assessment application	Contributes a validated interdisciplinary digital-assessment prototype integrating mathematics, science, reasoning, and an Islamic context; it provides feasibility rather than effectiveness evidence.
Hussein et al. (2019)	127 fifth-grade female students; quasi-experimental study	Critical thinking	Digital game-based learning using the ICMDCR framework	Reports significant improvement in critical thinking through an inquiry-, communication-, decision-, challenge-, and reward-based game, but no significant effect on motivation or science self-efficacy.
Ambarwati and Suyatna (2018)	Secondary students and physics teachers; developmental needs analysis and expert validation	Critical thinking and self-directed learning	Interactive electronic school book	Identifies design requirements for multimedia-supported independent physics learning; evidence is limited to needs analysis and expert validation.
Jaipal-Jamani (2024)	Preservice elementary teachers in Canada; mixed-methods one-group study	STEM problem-solving and metacognition	Educational robotics	Reports a small improvement in problem-solving and a large improvement in teaching self-efficacy; qualitative findings show expert-like problem-solving processes constrained by prior STEM knowledge.
Pujani et al. (2022)	23 science-education students; one-group pretest–posttest study	Problem-solving and generic science skills	Inquiry-based astronomy teaching materials	Reports moderate gains in problem-solving and generic science skills, although the lack of a comparison group limits effectiveness claims.
Dewi et al. (2023)	50 science-education students; research and development	Critical thinking and collaboration	Project-Based Laboratory Rotation Blended Learning	Provides evidence of product validity and moderate-to-high post-implementation skill profiles, but does not provide comparative evidence of improvement.
Nusantari et al. (2021)	Secondary students; research and development with limited- and large-scale trials	Critical thinking and metacognitive knowledge	Discovery learning combined with a metacognitive knowledge strategy	Shows that metacognitive prompts can strengthen discovery learning and support critical-thinking performance; the one-group implementation design requires cautious interpretation.
Koyunlu Ünlü and Dökme (2017)	447 preservice science teachers in Türkiye; descriptive survey	Critical-thinking disposition and epistemological beliefs	Personal epistemology framework	Identifies a moderate positive relationship between epistemological beliefs and critical-thinking disposition, contributing correlational

Study	Context and design	Primary HOT focus	Pedagogical, theoretical, or technological approach	Main contribution to the synthesis
				rather than intervention evidence.
Awang Noh et al. (2020)	Teachers and students in Malaysian science education; preliminary mixed-data study	HOTS requirements	Serious-game needs analysis informed by Bloom's revised taxonomy	Identifies instructional challenges, learner needs, and design requirements for serious games intended to support HOTS; it does not test learning effectiveness.
Marthaliakirana et al. (2022)	121 eleventh-grade biology students; three-group quasi-experimental study	Argumentation and critical thinking	Problem-based learning with metacognitive prompts and socioscientific issues	Provides comparatively strong evidence that metacognitive prompting within PBL supports argumentation and critical thinking more effectively than lower-intensity PBL conditions.
Cormas (2016)	27 preservice teachers; qualitative and process-oriented study	Problem-solving and epistemic reasoning	Sequenced mathematics–science integration and constructivist problem-solving	Shows that participants used shared mathematical and scientific problem-solving processes but retained disciplinary distinctions in how they interpreted the task.
Suardana et al. (2020)	34 science-education students; two-group quasi-experimental study	Critical thinking	Discovery learning with concept mapping or mind mapping	Finds no significant overall difference between the two mapping approaches, although differences emerged across specific critical-thinking indicators.
Mejía-Cáceres et al. (2023)	Preservice and in-service science teachers from Latin America; qualitative course-based study	Critical thinking and evaluation of evidence	Climate-change education, socioscientific issues, and critical discourse analysis	Illustrates how political speeches and climate discourse can be used to connect scientific, ethical, and political reasoning; findings are contextual and process-oriented.
Irawati and Sulisworo (2023)	57 seventh-grade students; quasi-experimental study	Critical thinking	Problem-based learning supported by IoT-based water-quality monitoring	Reports stronger critical-thinking outcomes for IoT-supported PBL than regular PBL and identifies learning interest as an important covariate.
Zuhaida (2018)	55 preservice science teachers; quasi-experimental study	Problem-solving, metacognition, and conceptual understanding	Metacognition-based basic-chemistry laboratory activities	Reports improved problem-solving performance and conceptual understanding through planning, monitoring, information management, and evaluation during laboratory work.

3.2 HOT Constructs Identified in the Included Studies

The synthesis identified four recurring HOT construct families: critical thinking and argumentation, problem-solving and scientific reasoning, creativity and innovation, and metacognition and self-regulation. Multiple coding was applied when a study addressed more than one construct.

Critical thinking and argumentation were the most prominent constructs. They were operationalized through interpretation, analysis, evaluation, inference, explanation, evidence assessment, reasoned judgment, and the justification of claims. These processes appeared in studies of problem-based learning, HOTS-oriented questioning, discovery learning, digital games, environmental investigation, and socioscientific learning (Amin et al., 2020; Hussein et al., 2019; Irawati & Sulisworo, 2023; Marthaliakirana et al., 2022; Mejía-Cáceres et al., 2023; Sidiq et al., 2021; Suardana et al., 2020).

Problem-solving and scientific reasoning were represented through problem identification, planning, information gathering, evidence interpretation, solution generation, testing, evaluation, and communication. These processes were examined through active learning, inquiry-based materials, educational robotics, mathematics-science integration, and laboratory activities (Cormas, 2016; Jaipal-Jamani, 2024; Kanphukiew & Nuangchalem, 2024; Pujani et al., 2022; Zuhaida, 2018).

Creativity and innovation were measured less consistently. Fathonah et al. (2024) explicitly assessed fluency, originality, flexibility, and elaboration, whereas other studies represented creativity through open-ended projects, alternative solutions, design activities, or product development without administering a dedicated creativity measure.

Metacognition and self-regulation appeared as both study outcomes and instructional mechanisms. The reported processes included planning, monitoring, information management, evaluation, reflection, and strategy revision. They were incorporated into online learning, problem-based learning, discovery learning, flipped classrooms, robotics, and laboratory instruction (Jaipal-Jamani, 2024; Maryani et al., 2022; Marthaliakirana et al., 2022; Nusantara et al., 2021; Suwardika et al., 2024; Zuhaida, 2018).

The coding distinguished construct emphasis from construct coverage. Construct emphasis referred to the HOT ability explicitly named and measured by the study authors, whereas construct coverage referred to additional cognitive or metacognitive processes represented in the learning activities but not independently assessed.

3.3 Frameworks Used in the Included Studies

The included studies employed cognitive taxonomies, learning and epistemological theories, competency frameworks, pedagogical models, instructional-design models, and assessment frameworks. These categories were coded separately according to their functions.

Bloom's revised taxonomy was the most visible cognitive taxonomy. It was used mainly to specify higher-level learning objectives, distinguish higher- from lower-order cognitive processes, and guide the construction of assessment items involving analysis, evaluation, and creation (Awang Noh et al., 2020; Bolat & Karamustafaoğlu, 2024; Hussein et al., 2019; Sidiq et al., 2021; Zulfiani et al., 2023).

Constructivist principles were reflected in inquiry-based, problem-based, discovery, project-based, active, and technology-supported learning. These approaches structured activities involving investigation, collaboration, evidence interpretation, and problem-solving (Amin et al., 2020; Hussein et al., 2019; Jaipal-Jamani, 2024; Kanphukiew & Nuangchalem, 2024; Pujani et al., 2022).

Metacognitive frameworks guided planning, monitoring, reflection, evaluation, and strategy revision in online learning, PBL, discovery learning, and laboratory activities (Maryani et al., 2022; Marthaliakirana et al., 2022; Nusantara et al., 2021; Zuhaida, 2018). The 21st Century Skills perspective was used to connect critical thinking and problem-solving with communication, collaboration, creativity, and self-regulation, although the included studies generally measured selected competencies rather than applying one standardized framework.

None of the 24 included studies operationalized the SOLO Taxonomy as its primary empirical framework. SOLO was therefore not coded as a dominant framework within the included corpus.

3.4 Methodological Characteristics and MMAT Findings

The study-level MMAT judgments reported in Appendix A were independently verified by two reviewers. Disagreements concerning research-design classification or individual appraisal criteria were resolved through discussion and reference to the MMAT guidance.

The comparative intervention evidence included studies of digital game-based learning, HOTS-question habituation, metacognition-integrated online learning, metacognitive PBL, multirepresentational creative problem-

solving, and IoT-supported PBL (Fathonah et al., 2024; Hussein et al., 2019; Irawati & Sulisworo, 2023; Marthaliakirana et al., 2022; Maryani et al., 2022; Sidiq et al., 2021). These studies commonly used non-randomized designs and context-specific samples.

One-group and pre-experimental studies reported changes after active-learning, inquiry, or developmental interventions but did not include equivalent comparison groups (Kanphukiew & Nuangchalerm, 2024; Pujani et al., 2022). Developmental studies contributed evidence on product validity, design requirements, feasibility, or implementation rather than comparative effectiveness (Ambarwati & Suyatna, 2018; Awang Noh et al., 2020; Dewi et al., 2023; Zulfiani et al., 2023).

Recurring methodological concerns identified through MMAT included non-randomized allocation, small or context-specific samples, limited control of confounding variables, one-group designs, incomplete reporting of allocation or outcome procedures, short follow-up periods, and uneven alignment between the stated HOT construct and the measurement instrument. No overall numerical quality score was calculated. The criterion-level judgments are presented in Appendix A.

3.5 Applications Across Educational Contexts

At the primary level, the included studies used scaffolded questioning, HOTS-oriented science problems, digital games, and interactive activities to assess or develop critical thinking and evidence-based reasoning (Hussein et al., 2019; Kinoshita, 2022; Sidiq et al., 2021).

Secondary-level studies applied problem-based learning, metacognitive prompting, environmental investigation, socioscientific issues, HOTS-oriented questions, and digital monitoring technologies. The assessed outcomes included critical thinking, argumentation, scientific problem-solving, environmental attitudes, and learning achievement (Amin et al., 2020; Irawati & Sulisworo, 2023; Marthaliakirana et al., 2022).

Tertiary and teacher-education studies examined educational robotics, online metacognitive learning, astronomy inquiry, blended laboratories, concept and mind mapping, chemistry experimentation, teacher-designed HOTS questions, and interdisciplinary digital assessment (Bolat & Karamustafaoğlu, 2024; Jaipal-Jamani, 2024; Maryani et al., 2022; Pujani et al., 2022; Suardana et al., 2020; Suwardika et al., 2024; Zulfiani et al., 2023; Zuhaida, 2018).

Teacher-focused studies also examined classroom interpretations of critical thinking, competence in developing HOTS-oriented questions, epistemological beliefs, and self-efficacy for STEM instruction (Bolat & Karamustafaoğlu, 2024; Jaipal-Jamani, 2024; Kinoshita, 2022; Koyunlu Ünlü & Dökme, 2017).

3.6 Technology-Supported Applications

Technology-supported studies used digital games, IoT-based environmental monitoring, educational robotics, flipped classrooms, online learning, interactive electronic materials, and digital assessment applications. These technologies were associated with activities involving inquiry, data interpretation, decision-making, feedback, reflection, and iterative problem-solving (Hussein et al., 2019; Irawati & Sulisworo, 2023; Jaipal-Jamani, 2024; Suwardika et al., 2024; Zulfiani et al., 2023).

The digital game developed by Hussein et al. (2019) incorporated inquiry, communication, decision-making, challenge, and rewards. Irawati and Sulisworo (2023) combined PBL with IoT-based water-quality monitoring, while Jaipal-Jamani (2024) used educational robotics to examine STEM problem-solving and teaching self-efficacy. Suwardika et al. (2024) integrated whiteboard animation and learning modules into a flipped classroom, and Zulfiani et al. (2023) developed a digital interdisciplinary HOTS assessment application.

The included corpus did not provide sufficient empirical evidence to identify augmented or virtual reality as dominant HOT technologies. AR- and VR-related applications were therefore not treated as established findings of the included studies.

3.7 Assessment of HOT

The included studies used multiple-choice and essay tests, open-ended problems, performance tasks, rubrics, observation sheets, questionnaires, self-report inventories, and researcher-developed disciplinary instruments.

Critical-thinking instruments assessed processes such as analysis, interpretation, evaluation, inference, explanation, deduction, and decision-making. Problem-solving instruments assessed problem identification, planning, implementation, evidence interpretation, solution evaluation, and communication. Creativity was measured through fluency, originality, flexibility, and elaboration, while metacognition and self-regulation were assessed through questionnaires, reflective activities, or indicators embedded in learning tasks.

Several studies reported expert validation, content-validity evidence, inter-rater agreement, or internal-consistency estimates. However, construct-instrument alignment varied across the corpus. Some studies used a broad HOTS label while measuring a narrower construct such as critical thinking, problem-solving, or cognitive achievement. The synthesis therefore classified findings according to the construct directly measured. Detailed instrument information and appraisal judgments are provided in Appendix A.

4. Discussion

This review synthesized 24 empirical studies on Higher-Order Thinking (HOT) in STEM and science education. Table 2 summarizes the included studies, while Appendix A provides the complete extraction matrix and study-level MMAT assessment. The appraisal findings were used to qualify the strength of interpretation rather than to exclude eligible studies. Accordingly, the discussion distinguishes comparative intervention evidence from descriptive, qualitative, developmental, feasibility, and needs-analysis evidence.

4.1 Active Learning and HOT Development

Across the corpus, active learning was consistently associated with opportunities to develop HOT. Problem-based learning, inquiry, discovery, project-based learning, and metacognitively supported instruction required learners to identify problems, interpret evidence, compare alternatives, justify conclusions, and reflect on their reasoning. Comparative studies generally reported favorable outcomes in critical thinking, argumentation, problem-solving, or creativity (Amin et al., 2020; Fathonah et al., 2024; Irawati & Sulisworo, 2023; Marthaliakirana et al., 2022; Maryani et al., 2022; Sidiq et al., 2021). One-group studies also reported improvements, but their designs could not exclude maturation, testing, or contextual effects (Kanphukiew & Nuangchalerm, 2024; Pujani et al., 2022).

These findings are consistent with external literature that identifies engagement, inquiry, and purposeful problem-solving as important conditions for advanced reasoning (Khurma & El Zein, 2024). Nevertheless, the MMAT assessment showed that much of the evidence was non-randomized, short-term, and context-specific. Active learning should therefore be interpreted as a promising condition for HOT development rather than as a universally effective intervention. Its influence depends on the quality of the task, the degree of learner autonomy, the available scaffolding, and the alignment between the intended construct and its assessment.

4.2 Disciplinary and Contextual Interpretation

Within STEM and science education, HOT was expressed primarily through scientific reasoning, problem-solving, evidence evaluation, argumentation, and the application of disciplinary knowledge to unfamiliar situations. Studies using inquiry, experimentation, robotics, integrated mathematics-science learning, and multirepresentational problem-solving required learners to analyze data, test explanations, and justify solutions (Amin et al., 2020; Cormas, 2016; Fathonah et al., 2024; Jaipal-Jamani, 2024; Pujani et al., 2022; Sidiq et al., 2021; Zuhaida, 2018).

However, the evidence also shows that HOT in science education is not limited to technical reasoning. Socioscientific and environmental tasks required learners to evaluate ethical consequences, political claims, and the credibility of competing evidence. Mejía-Cáceres et al. (2023) connected climate-science learning with analysis of political discourse, while Marthaliakirana et al. (2022) used socioscientific issues to strengthen evidence-based argumentation. These studies suggest that scientific HOT can include civic and ethical judgment when learning problems are socially situated.

The review does not support direct comparison between STEM and the humanities because the corpus was intentionally restricted to STEM and science education. Claims that humanities disciplines systematically emphasize ethical or qualitative reasoning therefore remain external interpretations rather than findings from the included

studies (Pouw et al., 2016; Tractenberg et al., 2019). The more defensible conclusion is that disciplinary context shapes the form through which HOT is demonstrated.

Regional interpretation also requires caution. The studies were conducted in different national and institutional settings, but the corpus was not designed for causal cross-country comparison. Nevertheless, curriculum demands, assessment practices, teacher preparation, technological access, and institutional support appeared to shape implementation. Teachers' interpretations of critical thinking influenced classroom practice (Kinoshita, 2022), prior STEM knowledge affected robotics-based problem-solving (Jaipal-Jamani, 2024), and curriculum alignment and teacher confidence influenced serious-game design requirements (Awang Noh et al., 2020). Political and ideological conditions also shaped how climate-related reasoning was addressed (Mejía-Cáceres et al., 2023).

These findings do not justify broad claims that particular cultures rely predominantly on rote learning. HOT implementation is better understood as the result of interactions among policy, assessment, teacher competence, resources, and locally situated beliefs about learning. Figure 2 therefore presents policy, cultural norms, framework variability, and differing interpretations of HOT as interacting influences rather than as quantified causal factors.

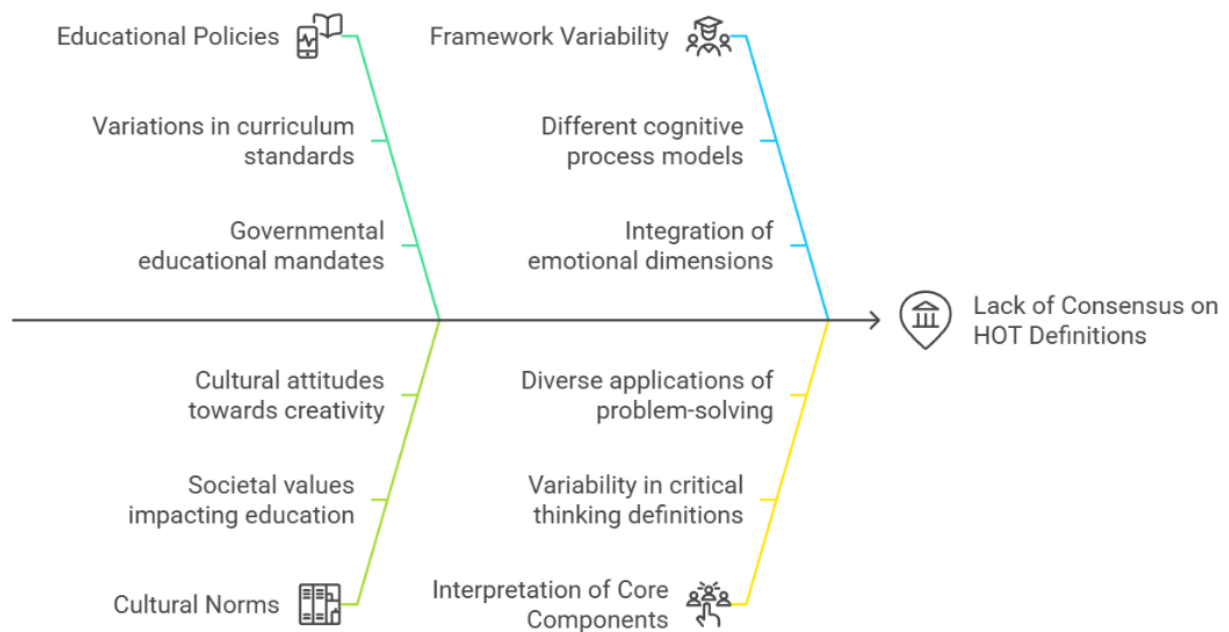


Figure 2. Conceptual synthesis of factors contributing to variability and the lack of consensus in HOT definitions and frameworks

4.3 Conceptual and Framework Implications

A central contribution of this review is the clarification that HOT is multidimensional. The four recurring construct families identified in the synthesis—critical thinking and argumentation, problem-solving and scientific reasoning, creativity and innovation, and metacognition and self-regulation—overlapped in practice but were not conceptually interchangeable. This distinction matters because an intervention may engage several HOT processes while measuring only one of them.

External psychometric research supports this multidimensional interpretation. The Mathematical Higher-Order Thinking Scale separates critical thinking, creative thinking, problem-solving, and metacognitive skills, while SHOTAPS distinguishes scientific creativity from scientific critical thinking (Shou, 2025; Zhou et al., 2024). These instruments illustrate why HOT should not automatically be represented through one undifferentiated score.

The included studies also demonstrated that frameworks serve different functions. Bloom's revised taxonomy was mainly used to specify cognitive targets and guide assessment design (Awang Noh et al., 2020; Bolat & Karamustafaoğlu, 2024; Hussein et al., 2019; Sidiq et al., 2021; Zulfiani et al., 2023). Constructivist, inquiry-based, problem-based, and discovery-learning models structured learning activities, while metacognitive frameworks

explained how learners planned, monitored, evaluated, and revised their thinking. The 21st Century Skills perspective connected cognition with communication, collaboration, creativity, and self-regulation. Treating these structures as equivalent would obscure whether a framework is being used to classify cognition, organize instruction, explain learning, or assess performance.

No included study operationalized the SOLO Taxonomy as its primary empirical framework. SOLO should therefore be retained only as an external theoretical comparison for the depth and integration of learning outcomes, not as a dominant finding from the 24-study corpus. Bloom and SOLO may be complementary, but they address different questions: Bloom classifies cognitive processes, whereas SOLO describes the structural complexity of demonstrated understanding.

Construct-instrument alignment remains a major concern. Several studies used the broad label HOTS while measuring a narrower construct such as critical thinking, problem-solving, or achievement. Others inferred creativity or metacognition from open-ended activities without using dedicated measures. The distinction between construct emphasis and construct coverage therefore remains essential. No defensible numerical estimate of label-construct concordance can be calculated because the included studies did not apply a common validation procedure.

Future measurement should begin with an explicit construct definition and then select tasks and instruments that provide corresponding evidence. Suitable approaches include exploratory and confirmatory factor analysis, Rasch analysis, content-validity indices, inter-rater reliability, and performance-based rubrics (Amalina & Vidákovich, 2022; Caratozzolo et al., 2022; Shou, 2025; Zhou et al., 2024).

4.4 Technology-Supported HOT

The findings indicate that technology supports HOT only when its features are linked to specific cognitive or metacognitive processes. Digital games, IoT-based investigation, educational robotics, flipped learning, online metacognitive instruction, interactive materials, and digital assessment created opportunities for inquiry, feedback, data interpretation, planning, and solution revision.

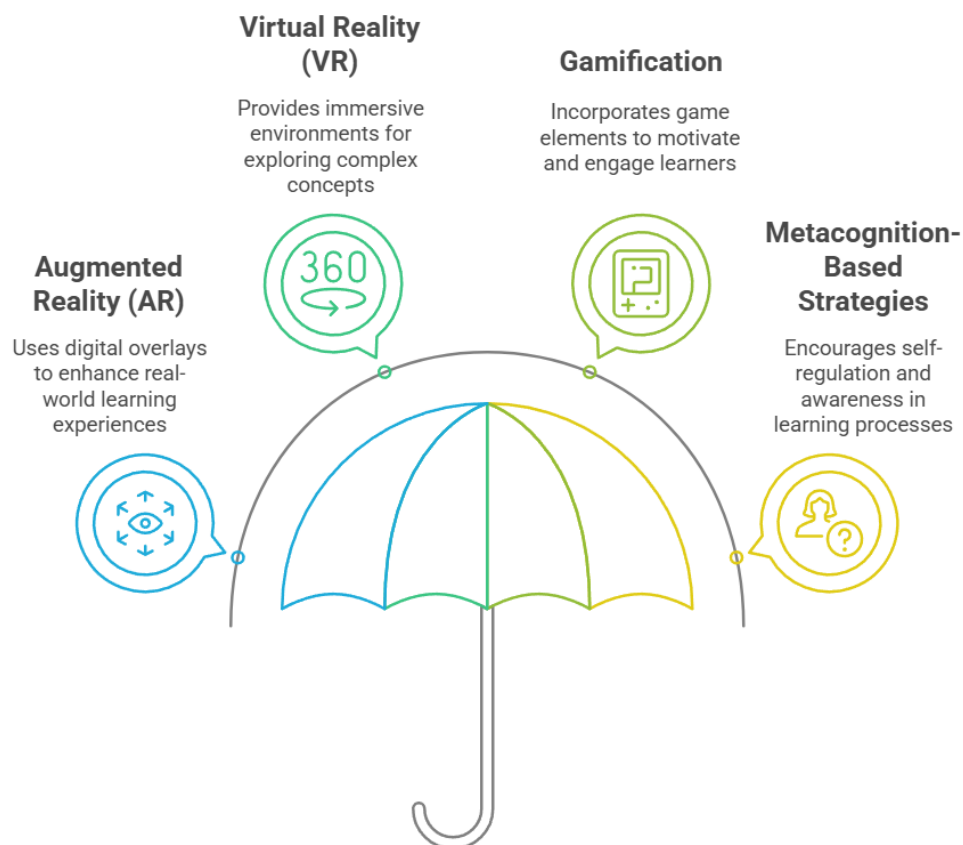


Figure 3. Technology-supported pathways and emerging opportunities for developing higher-order thinking

Digital game-based learning supported critical thinking when game mechanics were connected to inquiry, communication, decision-making, challenge, and feedback (Hussein et al., 2019). However, serious-game needs analysis provided evidence of design requirements rather than learning effectiveness (Awang Noh et al., 2020). IoT-supported PBL enabled learners to interpret authentic water-quality data and make evidence-based judgments, with learning interest operating as an important covariate (Irawati & Sulisworo, 2023). Educational robotics supported iterative planning and testing, although prior STEM knowledge constrained the range of strategies considered by preservice teachers (Jaipal-Jamani, 2024).

Flipped and online learning also supported HOT when they incorporated reflection and self-regulation rather than merely changing the delivery medium. Suwardika et al. (2024) combined whiteboard animation and modules with flipped learning, while Maryani et al. (2022) integrated planning, monitoring, and evaluation into synchronous and asynchronous science instruction. These findings suggest that pedagogical design, rather than technological novelty, determines the educational contribution of a tool.

The corpus did not provide sufficient HOT-specific evidence to establish augmented reality or virtual reality as effective interventions. AR and VR should therefore be presented as emerging research directions. Future studies should identify which immersive features are expected to support particular HOT processes and measure those processes directly. Figure 3 reflects this distinction by combining approaches supported in the included studies with prospective opportunities requiring stronger evidence.

4.5 Practical Implications

The main practical implication is the need for alignment among the intended HOT construct, the function of the selected framework, the learning task, and the assessment method. Educators should first specify whether they are targeting critical thinking, argumentation, problem-solving, creativity, metacognition, or a justified combination of these abilities. They should then select a framework that serves the required purpose rather than treating all taxonomies, pedagogies, and assessment systems as interchangeable.

Learning activities should make the targeted thinking process observable. Critical-thinking tasks should require evaluation of evidence and justification of conclusions. Problem-solving tasks should include planning, implementation, monitoring, and revision. Creativity tasks should permit multiple defensible responses and assess originality or flexibility. Metacognitive tasks should require learners to explain how they selected, monitored, and revised their strategies.

Assessment should mirror these demands. General achievement tests cannot automatically establish creativity, metacognitive regulation, or authentic problem-solving. Complex HOT outcomes may require combinations of validated tests, open-ended tasks, reflective records, observations, and analytic rubrics. Performance-based STEM assessments demonstrate how intended creativity and critical-thinking outcomes can be connected to observable design decisions and solution quality (Amalina & Vidákovich, 2022; Caratozzolo et al., 2022).

Teacher professional development should address this full alignment process. Teachers need support in defining constructs, designing cognitively demanding tasks, facilitating evidence-based discussion, developing scoring criteria, interpreting student reasoning, and providing feedback that promotes revision. The included studies indicate that teachers' knowledge, beliefs, prior experience, and assessment competence influence how HOT is implemented (Bolat & Karamustafaoğlu, 2024; Jaipal-Jamani, 2024; Kinoshita, 2022; Koyunlu Ünlü & Dökme, 2017).

The MMAT findings also indicate that practitioners should not treat all interventions as equally established. Comparative studies with clearer outcome measures provide stronger evidence than one-group, developmental, validation, feasibility, or needs-analysis studies. Appendix A enables readers to examine these differences at the study level.

4.6 Research Gaps and Interpretive Limitations

The included studies primarily assessed immediate or short-term outcomes. Few examined whether HOT gains were retained or transferred to new tasks and contexts. Future research should include delayed posttests and repeated measurements to distinguish temporary performance gains from sustained development. Long-term STEM studies demonstrate that multi-year interventions, robotics-program follow-up, and longitudinal cohort analysis are

feasible (Loof et al., 2022; Meschede et al., 2022; Remington et al., 2023), but such designs must include validated HOT-specific outcomes to demonstrate retention.

Transfer also remains unresolved. Improvement on an intervention-aligned test does not establish that learners can apply the same ability to another discipline or an authentic decision-making problem. Future studies should distinguish near transfer from far transfer and examine whether learners can recognize when and how to apply a strategy without explicit prompting.

The evidence base was geographically and methodologically uneven. Many studies used small, purposive, or institution-specific samples, limiting transferability. The corpus was also restricted to STEM and science education, preventing broader conclusions about HOT in the humanities, social sciences, vocational education, or professional learning. Future cross-context research should retain precise construct definitions and comparable measurement procedures.

The review itself also has limitations. The search was restricted to Scopus, English-language articles, and full conference papers published from 2014 to 2024. Relevant studies indexed elsewhere, published in other languages, or using adjacent terminology may have been missed. Heterogeneity in design, sample, intervention, duration, and measurement prevented statistical pooling. In addition, the MMAT assessment identified recurring concerns related to non-randomized allocation, one-group designs, limited control of confounding variables, incomplete reporting, short follow-up periods, and uneven construct-instrument alignment.

Accordingly, the findings should be interpreted as recurring patterns across a heterogeneous evidence base rather than as pooled estimates of effectiveness. The strongest conclusion is that HOT development depends less on adopting a particular framework or technology than on achieving coherence among construct definition, instructional mechanism, task design, and assessment.

5. Conclusion

This systematic review synthesized 24 empirical studies on Higher-Order Thinking (HOT) in STEM and science education. Table 2 summarizes the included studies, while Appendix A presents the full extraction matrix and study-level MMAT assessment. All studies were retained, and methodological appraisal was used to qualify the strength of interpretation rather than to determine eligibility.

The findings show that HOT is a multidimensional construct comprising four recurring families: critical thinking and argumentation, problem-solving and scientific reasoning, creativity and innovation, and metacognition and self-regulation. Their emphasis varied across educational levels, with primary education focusing on scaffolded inquiry, secondary education on interdisciplinary and socioscientific problem-solving, and tertiary education on authentic and professionally oriented applications.

Frameworks served different purposes. Bloom's revised taxonomy mainly guided cognitive objectives and assessment, constructivist and inquiry-oriented models structured learning activities, metacognitive frameworks supported self-regulation, and 21st Century Skills frameworks connected cognition with communication and collaboration. None of the included studies used SOLO as a primary empirical framework; therefore, it should be treated as a theoretical comparison rather than a dominant finding.

Active-learning and technology-supported approaches were generally associated with favorable HOT outcomes, particularly PBL, inquiry, metacognitive prompting, digital games, IoT-supported investigation, robotics, flipped learning, and digital assessment. However, many studies used non-randomized, short-term, or context-specific designs, limiting strong causal and generalizable conclusions. Evidence for AR and VR was insufficient to establish them as effective HOT interventions within the included corpus.

Overall, the review clarifies that effective HOT implementation depends on precise construct definitions, coherent alignment among frameworks, tasks, and assessments, valid measurement, teacher capacity, and contextual support. Its conclusions should be interpreted as a narrative and thematic synthesis rather than as pooled estimates of effectiveness.

6. Recommendations

Educational policy should define the HOT constructs expected at each educational level and align curriculum standards, instructional time, teacher development, and assessment requirements with those constructs. Policies

should move beyond broad commitments to HOT by specifying how analysis, evaluation, creativity, problem-solving, and metacognitive regulation will be taught and assessed.

Educational institutions should align the intended HOT construct with the framework, learning task, and assessment method. Critical thinking should be assessed through evidence evaluation and justification, problem-solving through planning and solution revision, creativity through originality and flexibility, and metacognition through planning, monitoring, and reflection. General achievement tests should not be treated as measures of the entire HOT domain.

Teacher professional development should focus on designing cognitively demanding tasks, facilitating evidence-based discussion, developing analytic rubrics, interpreting student reasoning, and integrating technology purposefully. Digital tools should be selected because their features support specific HOT processes, not merely because they are novel or engaging.

Future research should use stronger comparative and longitudinal designs, validated construct-specific instruments, delayed posttests, and transfer tasks. Studies should also distinguish product validity, feasibility, implementation, and effectiveness, while expanding HOT research beyond STEM and documenting contextual factors such as curriculum demands, assessment regimes, resources, and teacher preparation.

Author contribution

Muhali: conceptualization, methodology, data curation, formal analysis, and writing the original draft. **Hunaepi:** investigation, validation, data analysis, and manuscript review and editing. **Ni Nyoman Sri Putu Verawati:** supervision, methodological validation, interpretation of findings, and manuscript review and editing.

Conflict of Interest

There are no competing interests.

Funding

Not applicable.

Ethical statement

All included studies were derived from publicly available, peer-reviewed publications, and no primary data collection involving human participants was undertaken. Thus, no ethical approval or informed consent was required for this research.

Data availability

Not applicable.

Utilization of AI statement

During the preparation of this work, the authors used ChatGPT to enhance the clarity of the writing. After using the ChatGPT, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- Afikah, A., Astuti, S. R. D., Suyanta, S., Jumadi, J., & Rohaeti, E. (2022). Mobile learning in science education to improve higher-order thinking skills (HOTS) and communication skills: A systematic review. *International Journal of Advanced Computer Science and Applications*, 13(7), 698–704. <https://doi.org/10.14569/IJACSA.2022.0130782>
- Agustian, H. Y., Finne, L. T., Jørgensen, J. T., Pedersen, M. I., Christiansen, F. V., Gammelgaard, B., & Nielsen, J. A. (2022). Learning outcomes of university chemistry teaching in laboratories: A systematic review of empirical literature. *Review of Education*, 10(2). <https://doi.org/10.1002/rev3.3360>
- Amalina, I. K., & Vidákovich, T. (2022). An integrated STEM-based mathematical problem-solving test: Developing and reporting psychometric evidence. *Journal on Mathematics Education*, 13(4), 587–604. <https://doi.org/10.22342/jme.v13i4.pp587-604>

- Ambarwati, D., & Suyatna, A. (2018). Interactive design for self-study and developing students' critical thinking skills in electromagnetic radiation topic. *Journal of Physics: Conference Series*, 948, 012039. <https://doi.org/10.1088/1742-6596/948/1/012039>
- Amin, S., Utaya, S., Bachri, S., Sumarmi, S., & Susilo, S. (2020). Effect of problem-based learning on critical thinking skill and environmental attitude. *Journal for the Education of Gifted Young Scientists*, 8(2), 743–755. <https://doi.org/10.17478/jegys.650344>
- Awang Noh, S. N., Mat Zin, N. A., & Mohamed, H. (2020). Serious games requirements for higher-order thinking skills in science education. *International Journal of Advanced Computer Science and Applications*, 11(6), 211–216. <https://doi.org/10.14569/IJACSA.2020.0110627>
- Bolat, A., & Karamustafaoglu, S. (2024). Investigation of the items prepared by teachers who received training on item preparation measuring higher-order thinking skills. *Participatory Educational Research*, 11(6), 189–215. <https://doi.org/10.17275/per.24.86.11.6>
- Caratozzolo, P., Lara-Prieto, V., Hosseini, S., & Membrillo-Hernández, J. (2022). The use of video essays and podcasts to enhance creativity and critical thinking in engineering. *International Journal on Interactive Design and Manufacturing*, 16(3), 1231–1251. <https://doi.org/10.1007/s12008-022-00952-8>
- Chang, H.-Y., Liang, J.-C., & Tsai, C.-C. (2020). Students' context-specific epistemic justifications, prior knowledge, engagement, and socioscientific reasoning in a mobile augmented reality learning environment. *Journal of Science Education and Technology*, 29(3), 399–408. <https://doi.org/10.1007/s10956-020-09825-9>
- Chistyakov, A. A., Zhdanov, S. P., Avdeeva, E. L., Dyadichenko, E. A., Kunitsyna, M. L., & Yagudina, R. I. (2023). Exploring the characteristics and effectiveness of project-based learning for science and STEAM education. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(5), em2256. <https://doi.org/10.29333/ejmste/13128>
- Chung, E. Y. (2019). Facilitating learning of community-based rehabilitation through problem-based learning in higher education. *BMC Medical Education*, 19(1), 433. <https://doi.org/10.1186/s12909-019-1868-4>
- Cormas, P. C. (2016). Sequenced integration and the identification of a problem-solving approach through a learning process. *Eurasia Journal of Mathematics, Science and Technology Education*, 12(9), 2557–2574. <https://doi.org/10.12973/eurasia.2016.1258a>
- Cumming, M. M., Bettini, E., & Chow, J. C. (2023). High-quality systematic literature reviews in special education: Promoting coherence, contextualization, generativity, and transparency. *Exceptional Children*, 89(4), 412–431. <https://doi.org/10.1177/00144029221146576>
- Darmaji, D., Astalini, A., Kamid, K., & Triani, E. (2024). The influence of generic skills on gender differences in scientific processing and critical thinking in physics motion dynamics. *Jurnal Ilmiah Ilmu Terapan Universitas Jambi*, 8(2), 524–537. <https://doi.org/10.22437/jiituj.v8i2.37237>
- Darwis, R. H., Alimuddin, A., & Patimbangi, A. (2024). Higher-order thinking and critical thinking skills in problem-based learning environments: A systematic review. *Journal of Learning and Development Studies*, 4(2), 21–33. <https://doi.org/10.32996/jlds.2024.2.2.3>
- Dewi, N. R., Listiaji, P., Akhlis, I., Kurniawan, I. O., & Widyaningrum, R. A. (2023). Project-based laboratory rotation blended learning model to train students' critical thinking and collaboration in physics course. *E3S Web of Conferences*, 400, 01023. <https://doi.org/10.1051/e3sconf/202340001023>
- Doolittle, P., Wojdak, K., & Walters, A. (2023). Defining active learning: A restricted systematic review. *Teaching and Learning Inquiry*, 11. <https://doi.org/10.20343/teachlearninqu.11.25>
- Fathonah, S., Cahyono, E., Haryani, S., Sarwi, S., & Lestari, N. H. (2024). Application of multirepresentation-based creative problem-solving learning models to improve critical and creative thinking skills for students. *International Journal of Cognitive Research in Science, Engineering and Education*, 12(1), 185–200. <https://doi.org/10.23947/2334-8496-2024-12-1-185-200>
- Fleming, J. I., McLucas, A. S., & Cook, B. G. (2023). Review of four preregistration registries for special education researchers. *Remedial and Special Education*, 44(6), 495–505. <https://doi.org/10.1177/07419325231160293>
- Gard, A. M., Shariq, D., Albrecht, A., Lurie, A., Kim, H. C., Mitchell, C., & Hyde, L. W. (2025). A window into the state of the science: Current reporting practices related to generalizability in MRI and functional MRI studies. *Advances in Methods and Practices in Psychological Science*, 8(3). <https://doi.org/10.1177/25152459251372115>

- Gouvea, J. S. (2023). Integrating computation into science education. *CBE—Life Sciences Education*, 22(3), fe2. <https://doi.org/10.1187/cbe.23-05-0093>
- Hussein, M. H., Ow, S. H., Cheong, L. S., & Thong, M.-K. (2019). A digital game-based learning method to improve students' critical thinking skills in elementary science. *IEEE Access*, 7, 96309–96318. <https://doi.org/10.1109/ACCESS.2019.2929089>
- Irawati, F. I. M., & Sulisworo, D. (2023). Utilising smart water monitoring with IoT in science learning with problem-based learning model: Impact on critical thinking skills and the role of learning interest. *Journal of Pedagogical Research*, 7(4), 69–80. <https://doi.org/10.33902/JPR.202323708>
- Ishmail, H., & Ngene, N. C. (2024). Improving diversity in learning activities in a teaching curriculum in higher medical education. *Journal of Medical Education and Curricular Development*, 11. <https://doi.org/10.1177/23821205241293487>
- Itoi, K. (2024). Fostering inclusive learning and 21st-century skills: Creating translanguaging spaces in university content and language integrated learning courses. *International Journal of Applied Linguistics*, 36(1), 51–61. <https://doi.org/10.1111/ijal.12643>
- Jaipal-Jamani, K. (2024). Educational robotics and preservice teachers: STEM problem-solving skills and self-efficacy to teach. *Canadian Journal of Learning and Technology*, 50(4 Special Issue), 1–21. <https://doi.org/10.21432/cjlt28599>
- Jais, A., Yahaya, N., & Ibrahim, N. H. (2021). Sustainable development for teachers' questioning techniques for higher-order thinking skills (HOTS). *International Journal of Academic Research in Business and Social Sciences*, 11(1). <https://doi.org/10.6007/ijarbss/v11-i1/8163>
- Kanphukiew, S., & Nuangchalerm, P. (2024). Enhancing scientific problem-solving and learning achievement of lower secondary students through active learning. *Jurnal Pendidikan IPA Indonesia*, 13(1), 172–181. <https://doi.org/10.15294/jpii.v13i1.47672>
- Katende, E. N. (2023). Critical thinking in curriculum content: The case of Uganda. *Journal of Education and Practice*, 7(8), 1–18. <https://doi.org/10.47941/jep.1564>
- Khurma, O. A., & El Zein, F. (2024). Inquiry skills teaching and its relationship with UAE secondary school students' critical thinking: Systematic review of science teachers' perspectives. *Eurasia Journal of Mathematics, Science and Technology Education*, 20(2). <https://doi.org/10.29333/ejmste/14155>
- Kimmel, M., & van Alphen, F. (2025). The spectrum of distributed creativity: Tango dancing and its generative modalities. *Psychology of Aesthetics, Creativity, and the Arts*, 19(1), 112–132. <https://doi.org/10.1037/aca0000515>
- Kinoshita, H. (2022). Teaching of critical thinking skills by science teachers in Japanese primary schools. *Journal of Baltic Science Education*, 21(5), 801–816. <https://doi.org/10.33225/jbse/22.21.801>
- Koyunlu Ünlü, Z., & Dökme, I. (2017). Science teacher candidates' epistemological beliefs and critical thinking disposition. *Eurasian Journal of Educational Research*, 72, 203–220. <https://doi.org/10.14689/ejer.2017.72.11>
- Kožuh, I., Krajnc, R., Hadjileontiadis, L. J., & Debevc, M. (2018). Assessment of problem solving ability in novice programmers. *PLOS ONE*, 13(9), e0201919. <https://doi.org/10.1371/journal.pone.0201919>
- Liu, J., Ma, Y., Sun, X., Zhu, Z., & Xu, Y. (2021). A systematic review of higher-order thinking by visualizing its structure through HistCite and CiteSpace software. *The Asia-Pacific Education Researcher*, 31(6), 635–645. <https://doi.org/10.1007/s40299-021-00614-5>
- Liu, L. Q., & Mehigan, S. (2021). A systematic review of interventions used to enhance implementation of and compliance with the World Health Organization Surgical Safety Checklist in adult surgery. *AORN Journal*, 114(2), 159–170. <https://doi.org/10.1002/aorn.13469>
- Loof, H. D., Pauw, J. B., & Petegem, P. V. (2022). Integrated STEM education: The effects of a long-term intervention on students' cognitive performance. *European Journal of STEM Education*, 7(1), 13. <https://doi.org/10.20897/ejsteme/12738>
- López-Nicolás, R., López-López, J. A., Rubio-Aparicio, M., & Sánchez-Meca, J. (2021). A meta-review of transparency- and reproducibility-related reporting practices in published meta-analyses on clinical psychological interventions (2000–2020). *Behavior Research Methods*, 54(1), 334–349. <https://doi.org/10.3758/s13428-021-01644-z>

- Mak, H. W. F., Cheung, C. K., Lau, J., Chan, S. C. W., & Li, P. H. (2025). Transparency, reproducibility, and accessibility of clinical and experimental studies in allergy (TRACES): Study design and protocol. *Journal of Allergy and Clinical Immunology: Global*, 4(2), 100447. <https://doi.org/10.1016/j.jacig.2025.100447>
- Marthaliakirana, A. D., Suwono, H., Saefi, M., & Gofur, A. (2022). Problem-based learning with metacognitive prompts for enhancing argumentation and critical thinking of secondary school students. *Eurasia Journal of Mathematics, Science and Technology Education*, 18(9), em2148. <https://doi.org/10.29333/ejmste/12304>
- Maryani, I., Prasetyo, Z. K., Wilujeng, I., & Purwanti, S. (2022). Promoting higher-order thinking skills during online learning: The integration of metacognition in science for higher education. *International Journal of Evaluation and Research in Education*, 11(4), 1980–1988. <https://doi.org/10.11591/ijere.v11i4.23129>
- Mayo-Wilson, E., Grant, S., Supplee, L., Kianersi, S., Amin, A., DeHaven, A., & Mellor, D. T. (2021). Evaluating implementation of the Transparency and Openness Promotion (TOP) Guidelines: The TRUST process for rating journal policies, procedures, and practices. *Research Integrity and Peer Review*, 6(1). <https://doi.org/10.1186/s41073-021-00112-8>
- McMahon, K., Clark, I., Stensæth, K., Odell-Miller, H., Wosch, T., Bukowska, A. A., & Baker, F. A. (2022). Exploring shared musical experiences in dementia care: A worked example of a qualitative systematic review and thematic synthesis. *International Journal of Qualitative Methods*, 21. <https://doi.org/10.1177/16094069221127509>
- Mejía-Cáceres, M. A., Rieckmann, M., & Foleña Araújo, M. L. (2023). Political discourses as a resource for climate change education: Promoting critical thinking by closing the gap between science education and political education. *Sustainability*, 15(8). <https://doi.org/10.3390/su15086672>
- Meschede, T., Haque, Z., Warfield, M. E., Melchior, A., Burack, C., & Hoover, M. (2022). Transforming STEM outcomes: Results from a seven-year follow-up study of an after-school robotics program's impacts on freshman students. *School Science and Mathematics*, 122(7), 343–357. <https://doi.org/10.1111/ssm.12552>
- Müller, S., & Pietzner, V. (2020). Comparative study of divergent thinking among participants in a German science competition and pupils in a control group. *Eurasia Journal of Mathematics, Science and Technology Education*, 16(10), em1883. <https://doi.org/10.29333/ejmste/8408>
- Mutseekwa, C., & Maeresera, J. (2024). Evaluating contexts for implementing a competency-based curriculum in South and Eastern Africa: A meta-narrative review. *The Curriculum Journal*, 36(3), 368–383. <https://doi.org/10.1002/curj.301>
- Nusantari, E., Abdul, A., Damopolii, I., Alghafri, A. S. R., & Bakkar, B. S. (2021). Combination of discovery learning and metacognitive knowledge strategy to enhance students' critical thinking skills. *European Journal of Educational Research*, 10(4), 1781–1791. <https://doi.org/10.12973/EU-JER.10.4.1781>
- Parent, S., Caron, J., & Adams, G. (2022). Réflexion théorique et expérientielle sur le développement des compétences du 21e siècle des élèves en classe flexible au primaire. *Revue hybride de l'éducation*, 6(1), 57–91. <https://doi.org/10.1522/rhe.v6i1.1249>
- Pouw, W. T. J. L., van Gog, T., Zwaan, R. A., & Paas, F. (2016). Augmenting instructional animations with a body analogy to help children learn about physical systems. *Frontiers in Psychology*, 7, 860. <https://doi.org/10.3389/fpsyg.2016.00860>
- Pujani, N. M., Arsana, K. G. Y., Suma, K., Selamet, K., & Erlina, N. (2022). The effectiveness of introduction to astronomy teaching materials to improve problem-solving and generic science skills. *Jurnal Pendidikan IPA Indonesia*, 11(2), 333–340. <https://doi.org/10.15294/jpii.v11i2.34801>
- Pujawan, I. G. N., Rediani, N. N., Antara, I. G. W. S., Putri, N. N., & Bayu, G. W. (2022). Revised Bloom taxonomy-oriented learning activities to develop scientific literacy and creative thinking skills. *Jurnal Pendidikan IPA Indonesia*, 11(1), 47–60. <https://doi.org/10.15294/jpii.v11i1.34628>
- Rachmawati, D. L., & Purwati, O. (2021). EFL teachers' attitudes and competence in developing HOTS-based formative assessment. *Journal of English Educators Society*, 6(2). <https://doi.org/10.21070/jees.v6i2.1060>
- Redshaw, J. (2024). The recursive grammar of mental time travel. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1913). <https://doi.org/10.1098/rstb.2023.0412>
- Rehmat, A. P., & Hartley, K. (2020). Building engineering awareness: Problem-based learning approach for STEM integration. *Interdisciplinary Journal of Problem-Based Learning*, 14(1), 1–15. <https://doi.org/10.14434/ijpbl.v14i1.28636>

- Remington, T. F., Chou, P., & Topa, B. (2023). Experiential learning through STEM: Recent initiatives in the United States. *International Journal of Training and Development*, 27(3–4), 327–359. <https://doi.org/10.1111/ijtd.12302>
- Reyna, J., & Meier, P. (2020). Co-creation of knowledge using mobile technologies and digital media as pedagogical devices in undergraduate STEM education. *Research in Learning Technology*, 28. <https://doi.org/10.25304/rlt.v28.2356>
- Ruzaman, N. K., & Rosli, D. I. (2020). Inquiry-based education: Innovation in participatory inquiry paradigm. *International Journal of Emerging Technologies in Learning*, 15(10), 4–15. <https://doi.org/10.3991/ijet.v15i10.11460>
- Sandoval-Lentisco, A., Tortajada, M., López-Nicolás, R., López-López, J. A., Wagenmakers, E., Sánchez-Meca, J., & Hardwicke, T. E. (2025). Preregistration of psychology meta-analyses: A cross-sectional study of prevalence and practice. *Advances in Methods and Practices in Psychological Science*, 8(1). <https://doi.org/10.1177/25152459241300113>
- Sasson, I., & Tifferet, S. (2025). Active and passive explicit instruction in question-formulating in higher education. *The Curriculum Journal*. <https://doi.org/10.1002/curj.70004>
- Shou, X., Han, K., & Sheng-quan, L. (2025). Primary students' core dimension of higher-order thinking in science: Measurement and reflections. *Psychology in the Schools*, 62(11), 4760–4778. <https://doi.org/10.1002/pits.70035>
- Sidiq, Y., Ishartono, N., Dessty, A., Prayitno, H. J., Anif, S., & Hidayat, M. L. (2021). Improving elementary school students' critical thinking skill in science through HOTS-based science questions: A quasi-experimental study. *Jurnal Pendidikan IPA Indonesia*, 10(3), 378–386. <https://doi.org/10.15294/JPII.V10I3.30891>
- Smet, C. D., Raileanu, M.-B., & Roméro, M. (2021). Étude de la littérature sur la créativité en sciences de l'éducation dans les pays francophones. *McGill Journal of Education / Revue des sciences de l'éducation de McGill*, 55(3), 588–618. <https://doi.org/10.7202/1083424ar>
- Sotiriou, S. A., Lazoudis, A., & Bogner, F. X. (2020). Inquiry-based learning and e-learning: How to serve high and low achievers. *Smart Learning Environments*, 7(1), 29. <https://doi.org/10.1186/s40561-020-00130-x>
- Suardana, I. N., Redhana, I. W., & Yunitasari, N. P. M. (2020). Students' critical thinking skills comparison in discovery learning based on constructing concept mapping and mind mapping. *Journal of Physics: Conference Series*, 1521(4), 042089. <https://doi.org/10.1088/1742-6596/1521/4/042089>
- Suwardika, G., Sopandi, A. T., Indrawan, I. P. O., & Masakazu, K. (2024). A flipped classroom with whiteboard animation and modules to enhance students' self-regulation, critical thinking and communication skills: A conceptual framework and its implementation. *Asian Association of Open Universities Journal*, 19(2), 135–152. <https://doi.org/10.1108/AAOUJ-10-2023-0115>
- Tractenberg, R. E., Lindvall, J. M., Attwood, T. K., & Via, A. (2019). The Mastery Rubric for Bioinformatics: A tool to support design and evaluation of career-spanning education and training. *PLOS ONE*, 14(11), e0225256. <https://doi.org/10.1371/journal.pone.0225256>
- Tremblay, C., & Poëllhuber, B. (2022). Analyse qualitative de référentiels de compétences du XXI^e siècle, numériques et informationnelles: Tendances mondiales observées. *Formation et Profession*, 30(2), 1. <https://doi.org/10.18162/fp.2022.648>
- Triana, H., Abustang, P. B., Utomo, E., Rakhman, G. G. F., & Fahrurrozi, F. (2023). Assessment evaluation using SOLO taxonomy for measuring levels of critical thinking skills: PYP International Baccalaureate case study. *Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram*, 11(2), 581. <https://doi.org/10.33394/j-ps.v11i2.7894>
- Wilson, R., Hrubý, A., Perez-Zapata, D. I., van der Kleij, S. W., & Apperly, I. A. (2023). Is recursive “mindreading” really an exception to limitations on recursive thinking? *Journal of Experimental Psychology: General*, 152(5), 1454–1468. <https://doi.org/10.1037/xge0001322>
- Wrzesien, M., & Alcañiz Raya, M. (2010). Learning in serious virtual worlds: Evaluation of learning effectiveness and appeal to students in the E-Junior project. *Computers & Education*, 55(1), 178–187. <https://doi.org/10.1016/j.compedu.2010.01.003>
- Yanti, F. A., & Thohir, M. A. (2024). Higher-order thinking skills in science learning: A systematic review from 2014–2023. *International Journal of Evaluation and Research in Education*, 13(4), 2419–2427. <https://doi.org/10.11591/ijere.v13i4.28082>

- Zana, F. M., Sa'dijah, C., & Susiswo, S. (2024). The cognitive alignment of mathematics teachers' assessments and its curriculum. *International Journal of Evaluation and Research in Education*, 13(3), 1561. <https://doi.org/10.11591/ijere.v13i3.26814>
- Zhang, W., & Wang, Z. (2021). Theory and practice of VR/AR in K–12 science education: A systematic review. *Sustainability*, 13(22), 12646. <https://doi.org/10.3390/su132212646>
- Zhou, Y., Ning, Y., Chen, J., Zhang, W., & Wijaya, T. T. (2024). Development and validation of a mathematical higher-order thinking scale for high school students. *Psychology in the Schools*, 61(8), 3160–3192. <https://doi.org/10.1002/pits.23213>
- Zuhaida, A. (2018). Implementation of basic chemistry experiment based on metacognition to increase problem-solving and build concept understanding. *Journal of Physics: Conference Series*, 1006, 012041. <https://doi.org/10.1088/1742-6596/1006/1/012041>
- Zulfiani, Z., Suwarna, I. P., Muin, A., Mulyati, T., & El Islami, R. A. Z. (2023). Developing the MathSci 21st app: Enhancing higher-order thinking skills assessment in mathematics and science education within an Islamic context. *International Journal of Advanced and Applied Sciences*, 10(8), 19–31. <https://doi.org/10.21833/ijaas.2023.08.003>

Appendix A

Table A1. Comprehensive analysis and synthesis of the 24 studies included in the review

Study and publication context	Educational context and sample	Study design and MMAT category	HOT construct emphasis and coverage	Framework, pedagogy, and technology	Assessment and analysis	Main synthesis finding	MMAT appraisal and appropriate synthesis use
Kinoshita (2022) Journal article; Japan	Primary education / in-service teachers; Primary science education 291 primary-school teachers who taught science (136 male, 155 female; careers 1–38 years)	Cross-sectional questionnaire survey with ANOVA and SEM MMAT category: Quantitative descriptive	Emphasis: Critical-thinking teaching practices Coverage: Evidence use, data reliability, skepticism, multifaceted thinking, conclusion evaluation	Evidence-oriented critical-thinking teaching was operationalized explicitly; no primary Bloom/SOLO framework. Pedagogy/intervention: None; mapping of teachers' reported instructional practices Technology: None	New critical-thinking teaching questionnaire; factor/reliability analyses; ANOVA and structural equation modeling	Teachers emphasized conducting experiments but less often emphasized repeated trials, data reliability, and cautious conclusions; career length was associated with some practices.	Minor concerns. Core for teacher-practice/application synthesis; not effectiveness evidence
Bolat & Karamustafaoglu (2024) Journal article; Türkiye	Teacher professional development / in-service science teachers; Science assessment 8 volunteer science teachers selected for maximum variation from a prior cohort of 25	Exploratory special case study; artefact analysis using a 21-item checklist and three raters MMAT category: Quantitative descriptive	Emphasis: HOTS item-writing competence Coverage: Critical, analytical, creative thinking, problem-solving, decision-making	Bloom/Anderson–Krathwohl and Taba were explicitly used for HOTS item design. Pedagogy/intervention: Eight-week training, followed by item preparation over an extended period Technology: None	Researcher-developed 21-item checklist; three-rater coding; frequencies and percentages	After training, 43% of prepared items were judged to measure HOTS; strengths and weaknesses varied across criteria and grade levels.	Some concerns. Retain for assessment and teacher-capacity synthesis; do not use as student-effectiveness evidence
Amin et al. (2020) Journal article; Indonesia	Higher education; Environmental geography / sustainable development	Quasi-experimental pretest–posttest control-group design	Emphasis: Critical thinking Coverage: Problem formulation, argumentation, deduction,	PBL (Arends) and Ennis critical-thinking indicators were explicitly used. Pedagogy/inte	Validated essay Critical Thinking Skill Test (KR-20=0.756); environmen	PBL group showed higher critical-thinking gain and environmental-attitude outcomes than	Some concerns. Use cautiously in effectiveness synthesis; label environmental-geography

Study and publication context	Educational context and sample	Study design and MMAT category	HOT construct emphasis and coverage	Framework, pedagogy, and technology	Assessment and analysis	Main synthesis finding	MMAT appraisal and appropriate synthesis use
	50 university students (25 PBL; 25 control), purposively selected from two classes	MMAT category: Quantitative non-randomized	induction, evaluation, solution selection; environmental attitude	Intervention: PBL on sustainable environmental-development problems versus conventional learning Technology: None	Attitude questionnaire	the control group.	context explicitly
Suwardika et al. (2024) Journal article; Indonesia	Higher education / in-service elementary teachers; Elementary school science education 83 students/elementary teachers in three groups (28 conventional+module; 27 flipped+module; 28 flipped+whiteboard animation+module)	Mixed methods; non-equivalent three-group quasi-experiment plus observation, documents, and interviews MMAT category: Mixed methods	Emphasis: Critical thinking, self-regulation, communication Coverage: Analysis, evaluation, reflection, learning regulation, interpersonal communication	Flipped learning, Zimmerman self-regulation, Facione critical thinking, and 21st-century competencies were explicit. Pedagogy/intervention: Three webinar-tutorial conditions with modules and whiteboard animation Technology: Online webinar, modules, whiteboard animation	Validated self-regulation questionnaire, critical-thinking test/rubric, ICCS; observation, worksheet documentation, interviews; MANCOVA	The flipped condition with whiteboard animation and modules produced stronger multivariate outcomes than the comparison conditions.	Some concerns. Core intervention evidence with causal caution
Fathonah et al. (2024) Journal article; Indonesia	Higher education; Nutrition and health across science-related study programs 227 university students across four programs	R&D/ADDIE with non-randomized comparative pretest–posttest implementation MMAT category: Quantitative	Emphasis: Critical and creative thinking Coverage: Analysis, inference, evaluation, decision-making; fluency, flexibility, originality, elaboration	Creative Problem Solving, multirepresentation, and ADDIE were explicitly integrated. Pedagogy/intervention: Multirepresentation-based CPS on balanced-	Critical- and creative-thinking tests with reported construct validity/reliability; N-gain and Mann–Whitney tests	Both groups improved, but MBCPS yielded larger gains in critical and creative thinking.	Some concerns. Core effectiveness evidence with non-randomized-design caveat

Study and publication context	Educational context and sample	Study design and MMAT category	HOT construct emphasis and coverage	Framework, pedagogy, and technology	Assessment and analysis	Main synthesis finding	MMAT appraisal and appropriate synthesis use
	(114 MBCPS; 113 problem-solving comparison)	tive non-randomized		nutrition problems versus conventional problem-solving Technology: Multiple verbal, visual, and mathematical representations; no primary digital technology			
Maryani et al. (2022) Journal article; Indonesia	Higher education / preservice elementary teachers; Online science education Needs analysis: 21 lecturers; content validation: 7 experts; experiment: 119 students (41 experimental, 39 control 1, 39 control 2)	ADDIE R&D with a reported randomized pretest–posttest comparison on-group experiment MMAT category: Randomized controlled trial (allocation method unclear)	Emphasis: HOTS integrated with metacognition Coverage: Planning, monitoring, evaluation, reflection, problem-solving and decision-making	ADDIE, metacognitive regulation, constructivism, and ATC21S/21st-century skills were explicit. Pedagogy/intervention: Metacognition-integrated online science learning versus PBL and experiment-based controls Technology: Google Classroom, Google Meet, Google Forms, YouTube, PhET	Expert-validation forms; HOTS assessment; implementation observation; GLM/MANOVA	All groups improved, with the metacognition-integrated group showing the largest change; partial eta squared was interpreted as a large effect.	Some concerns. Core evidence, but avoid strong causal wording until randomization details are confirmed
Sidiq et al. (2021) Journal article; Indonesia	Primary education (Grade 5); Elementary science 60 students from one school (30 intervention; 30 control), classes	Quasi-experimental non-equivalent pretest–posttest control-group design MMAT	Emphasis: Critical thinking Coverage: Analysis, inference, evaluation, induction, deduction	Revised Bloom and critical-thinking indicators were explicitly aligned with HOTS questions. Pedagogy/intervention:	Three HOTS essay questions; expert review, Aiken's V, reliability=0.64; two assessors with	The intervention group had a significantly higher posttest mean than the control group.	Some concerns. Core evidence with limited generalizability

Study and publication context	Educational context and sample	Study design and MMAT category	HOT construct emphasis and coverage	Framework, pedagogy, and technology	Assessment and analysis	Main synthesis finding	MMAT appraisal and appropriate synthesis use
	selected by cluster random sampling	category: Quantitative non-randomized		Habituation with HOTS-based science essay questions versus lecture-based instruction Technology: None	Cohen's kappa		
Darmaji et al. (2024) Journal article; Indonesia	Secondary education; Physics: dynamics of rectilinear motion 180 high-school students from three schools (reported as 90 male and 90 female)	Comparative cross-sectional quantitative study MMAT category: Quantitative descriptive	Emphasis: Critical thinking, generic science skills, science-process skills Coverage: Observation, classification, communication, analysis, reasoning and problem-solving	Generic science-skills and science-process frameworks were explicit; the HOT framework was weakly operationalized. Pedagogy/intervention: No intervention; gender and skill comparisons/associations Technology: None	Observation sheets, five-item critical-thinking essay test, generic-skills questionnaire; ANOVA, t-tests, regression	The paper reports gender differences and relationships among generic skills, process skills, and critical thinking.	Major concerns. Use only for descriptive context after verifying tables; exclude from strong inferential claims
Kanphukiew & Nuangchalem (2024) Journal article; Thailand	Lower secondary education (Grade 8); Science: force and motion 30 eighth-grade students from one school	Pre-experimental one-shot/post test-only case design with criterion comparison MMAT category: Quantitative non-randomized	Emphasis: Scientific problem-solving Coverage: Critical thinking, decision-making, application and transfer	Constructivism and active learning were explicit; broader HOT alignment was inferred through problem-solving tasks. Pedagogy/intervention: Five active-science-learning plans over 15 hours/5 weeks Technology: Digital media used as part of activities, not	Achievement test, scientific problem-solving assessment, observation/reflection data, satisfaction questionnaire; one-sample t-tests	Post-intervention achievement and problem-solving scores exceeded preset criteria; satisfaction was high.	Major concerns. Use as preliminary/descriptive evidence only; do not infer causality

Study and publication context	Educational context and sample	Study design and MMAT category	HOT construct emphasis and coverage	Framework, pedagogy, and technology	Assessment and analysis	Main synthesis finding	MMAT appraisal and appropriate synthesis use
				isolated as an intervention			
Zulfiani et al. (2023) Journal article; Indonesia	Secondary education (Grade XI); Integrated mathematics, biology and physics assessment 32 students, 1 teacher, and 10 validators (5 material, 3 media, 2 language)	R&D using Akker's model; preliminary and prototype stages with limited trial MMAT category: Quantitative descriptive	Emphasis: HOTS assessment and interdisciplinary reasoning Coverage: Application, reasoning, problem-solving and cross-disciplinary integration	Bloom ladder analogy, TIMSS cognitive levels, and Akker R&D were explicitly used. Pedagogy/intervention: Development and limited trial of the Android-based MathSci 21st assessment app Technology: Android app, admin database, digital assessment	Expert-validation questionnaires, student/teacher responses, multiple-choice and true-false items; CVR/CVI and Aiken's V	The prototype was judged feasible/valid and practical for rapid assessment and feedback.	Some concerns. Use for technology/assessment design, not intervention effectiveness
Hussein et al. (2019) Journal article; Malaysia	Primary education (Grade 5); Elementary science 127 female students from four classes (62 game; 65 conventional)	Quasi-experimental class-based comparison with pretest and posttest MMAT category: Quantitative non-randomized	Emphasis: Critical thinking Coverage: Hypothesis identification, induction, deduction, explanation, argument evaluation, decision-making	The ICMDCR serious-game framework was explicitly used. Pedagogy/intervention: Ecoship Endeavour game versus conventional science teaching Technology: Role-playing educational computer game	Yeh critical-thinking measure (20 MCQ; alpha reported as 0.80), motivation and self-efficacy questionnaires	The game improved critical-thinking outcomes relative to conventional instruction, but not motivation or self-efficacy.	Some concerns. Core evidence with population and allocation caveats

Study and publication context	Educational context and sample	Study design and MMAT category	HOT construct emphasis and coverage	Framework, pedagogy, and technology	Assessment and analysis	Main synthesis finding	MMAT appraisal and appropriate synthesis use
Ambarwati & Suyatna (2018) Conference proceeding; Indonesia	Secondary education (Grade XII); Physics: electromagnetic radiation Needs analysis: 87 students and 2 physics teachers from two schools; expert validators not clearly enumerated	ADDIE R&D limited to needs analysis and expert validation of design MMAT category: Quantitative descriptive	Emphasis: Critical thinking and self-study Coverage: Analysis and problem-solving were intended but not directly measured	ADDIE was explicit; critical-thinking improvement was an intended design outcome rather than a measured effect. Pedagogy/intervention: Design of an interactive ESB; no classroom implementation of HOT outcomes Technology: Interactive electronic school book with images, animation, simulation, video, quizzes and feedback	Needs questionnaire; expert design-validation score sheet; descriptive means	Stakeholders reported a need for interactive materials, and experts judged design components suitable.	Major concerns. Use only for technology-design opportunity; not evidence of HOT improvement
Jaipal-Jamani (2024) Journal article; Canada	Higher education / preservice teachers; STEM/science teacher education 57 consented; 36 completed both pre/post instruments; two groups (2 and 3 participants) selected for qualitative analysis	Mixed methods; one-group pretest–posttest plus qualitative discourse analysis MMAT category: Mixed methods	Emphasis: STEM problem-solving and teaching self-efficacy Coverage: Goal setting, framing, planning, information interpretation, reflection, implications and communication	Price et al.'s STEM problem-solving model and Bandura self-efficacy were explicit. Pedagogy/intervention: Three-hour LEGO Mindstorms robotics and coding activity Technology: Educational robotics and coding	Problem-solving and self-efficacy pre/post questionnaires, worksheet, group-interaction transcripts	Self-efficacy improved with a large effect and problem-solving with a small effect; discourse showed expert-like processes constrained by prior STEM knowledge.	Some concerns. Use for process/mechanism and preliminary outcome evidence, not strong causal inference

Study and publication context	Educational context and sample	Study design and MMAT category	HOT construct emphasis and coverage	Framework, pedagogy, and technology	Assessment and analysis	Main synthesis finding	MMAT appraisal and appropriate synthesis use
Pujani et al. (2022) Journal article; Indonesia	Higher education / preservice science teachers; Astronomy / earth and space science 23 fourth-semester science-education students in one class	Quasi-experimental one-group pretest–posttest design MMAT category: Quantitative non-randomized	Emphasis: Problem-solving and generic science skills Coverage: Polya stages, observation, inference, causal reasoning, modeling and concept construction	Inquiry learning and Polya's problem-solving stages were explicit. Pedagogy/intervention: Inquiry-based introductory astronomy teaching materials and practical activities Technology: Teaching materials and astronomy practical work; no central digital tool	Ten essay problem-solving items; multiple-choice generic-science-skills test; paired t-tests and N-gain	Problem-solving and generic-science-skill scores increased from pretest to posttest.	Major concerns. Use as preliminary evidence only
Dewi et al. (2023) Conference proceeding; Indonesia	Higher education; Basic physics: projectile motion 50 science-education students; 5 expert validators	ADDIE R&D with expert validation and single-group post-implementation description MMAT category: Quantitative descriptive	Emphasis: Critical thinking and collaboration Coverage: Basic clarification, inference, advanced clarification, strategy/tactics	PjBL, laboratory-rotation blended learning, ADDIE, and 4Cs were explicitly integrated. Pedagogy/intervention: PjLRBL model using projects, virtual practicum and digital modules Technology: Virtual laboratory and web-based flipbook/digital module	Expert validation using Aiken's V; critical-thinking test; collaboration observation rubric; categorical descriptive analysis	Products were judged valid; after implementation, 48% of students were categorized as high in critical thinking and 56% as high in collaboration.	Major concerns. Use for design/application mapping, not effectiveness claims

Study and publication context	Educational context and sample	Study design and MMAT category	HOT construct emphasis and coverage	Framework, pedagogy, and technology	Assessment and analysis	Main synthesis finding	MMAT appraisal and appropriate synthesis use
Nusantari et al. (2021) Journal article; Indonesia	Secondary education (Grade 12); Biology/genetics 3 experts and 1 practitioner; 32 students in limited trial; 59 students in large-scale trial	4-D R&D with one-group pretest–posttest effectiveness trial MMAT category: Quantitative non-randomized	Emphasis: Critical thinking integrated with metacognitive knowledge Coverage: Assumption, inference, deduction, interpretation, evaluation; declarative, procedural and conditional knowledge	Discovery learning, metacognitive knowledge strategy, and the 4-D model were explicit. Pedagogy/intervention: Discovery-learning instructional tools with metacognitive prompts in genetics Technology: No central technology	Lesson-plan/student worksheet validation, implementation observation, student responses, critical-thinking test; related t-test and N-gain	Tools met validity/practicality criteria, and critical-thinking scores increased in the one-group large-scale trial.	Major concerns. Use for pedagogical design and preliminary effects; avoid causal language
Koyunlu Ünlü & Dökme (2017) Journal article; Türkiye	Higher education / preservice science teachers; Science teacher education 447 students (357 female, 90 male; 115 freshmen, 119 sophomores, 129 juniors, 84 seniors)	Cross-sectional descriptive survey and correlational study MMAT category: Quantitative descriptive	Emphasis: Critical-thinking disposition and epistemological beliefs Coverage: Analyticity, open-mindedness, inquisitiveness and epistemic sophistication	Schommer's epistemological-beliefs model and CCTDI were explicit. Pedagogy/intervention: None Technology: None	Scale of Epistemological Beliefs and CCTDI; Mann–Whitney U, Kolmogorov–Smirnov and correlation analyses	Female students had more developed epistemological beliefs; epistemological beliefs and critical-thinking dispositions were moderately positively correlated.	Minor concerns. Use to distinguish disposition from performance and to discuss epistemic foundations
Awang Noh et al. (2020) Journal article; Malaysia	Primary education and teacher-support context; Science education / educational game design 2 School Improvement Specialist Coaches, 8 science teachers, 60 primary	Preliminary mixed-method needs analysis: interviews, document analysis and questionnaire survey MMAT	Emphasis: HOTS requirements for serious games Coverage: Apply, analyze, evaluate, create; critical thinking, problem-solving, reflection and decision-making	Revised Bloom and serious-game design requirements were explicitly used. Pedagogy/intervention: No HOTS intervention; requirements elicitation for a future serious game Technology:	Semi-structured interviews, daily lesson-plan analysis, student interface questionnaire; descriptive means	Teachers reported difficulties implementing HOTS, lesson plans rarely specified HOTS elements, and students identified language/audio/animation design needs.	Some concerns. Use for technology-design barriers and requirements only

Study and publication context	Educational context and sample	Study design and MMAT category	HOT construct emphasis and coverage	Framework, pedagogy, and technology	Assessment and analysis	Main synthesis finding	MMAT appraisal and appropriate synthesis use
	students, and teachers' daily lesson plans	category: Mixed methods		Two existing educational games evaluated for interface requirements			
Marthalia kirana et al. (2022) Journal article; Indonesia	Secondary education (Grade 11); Biology and socioscientific issues 121 students in three intact groups: M-PBL 45; H-PBL 35; L-PBL 41	Quasi-experimental non-equivalent pretest-posttest three-group design MMAT category: Quantitative non-randomized	Emphasis: Argumentation and critical thinking Coverage: Claims, data, warrants, backing, rebuttal; assumptions, deduction, interpretation, inference and decision-making	PBL, metacognitive prompts, Toulmin argumentation, SSI, and social constructivism were explicit. Pedagogy/intervention: M-PBL versus high- and low-intensity PBL over three biology modules Technology: No central technology	Argumentation essay test and critical-thinking multiple-choice test; pre/post comparative analyses	M-PBL produced higher argumentation and critical-thinking outcomes than both comparison PBL conditions.	Minor concerns. Core evidence for framework-task-assessment alignment
Cormas (2016) Journal article; United States	Higher education / preservice elementary and special-education teachers; Integrated mathematics and science education 27 preservice teachers in two course sections	Constructivist grounded-theory study using constant comparison MMAT category: Qualitative	Emphasis: Problem-solving and epistemic understanding Coverage: Conjecture/testing, reasoning, experimentation, data collection, disciplinary categorization and reflection	Constructivism, inquiry/5E, NCTM problem-solving, and sequenced integration were explicit. Pedagogy/intervention: Fifteen-week integrated course and lever-simulator problem-solving activity Technology: Lever simulator	Student solutions and explanations, course artefacts and qualitative coding/constant comparison	Participants used similar mathematics/science problem-solving processes but usually categorized the lever problem as mathematics, revealing disciplinary epistemic distinctions.	Minor concerns. Core for conceptualization and process synthesis; not quantitative effectiveness evidence

Study and publication context	Educational context and sample	Study design and MMAT category	HOT construct emphasis and coverage	Framework, pedagogy, and technology	Assessment and analysis	Main synthesis finding	MMAT appraisal and appropriate synthesis use
Suardana et al. (2020) Conference proceeding; Indonesia	Higher education; Basic chemistry 34 science-education students (17 concept-mapping; 17 mind-mapping)	Quasi-experimental non-equivalent pretest-posttest two-group design MMAT category: Quantitative non-randomized	Emphasis: Critical thinking Coverage: Interpretation, analysis, evaluation, inference and explanation	Discovery learning, concept/mind mapping, and CCTST indicators were explicit. Pedagogy/intervention: Discovery learning with concept-map construction versus mind-map construction Technology: No central technology	Critical-thinking test based on California indicators; nonparametric baseline comparison and posttest analyses	Overall critical-thinking scores did not differ significantly between groups, although indicator-level patterns differed.	Some concerns. Core for balanced outcome synthesis; do not omit the null finding
Mejía-Cáceres et al. (2023) Journal article; Brazil / Latin America	Preservice and in-service science-teacher education; Climate-change and sustainability education Two presidential speeches as primary documents; online course with 33 registrants and 16 full participants from six Latin-American countries	Critical qualitative research and critical discourse analysis embedded in an educational-course case MMAT category: Qualitative	Emphasis: Critical thinking in climate-change education Coverage: Evaluation of discourse and evidence, ideology/power analysis, policy reasoning and transformative reflection	Freirean education, SSI, nature of science, and critical discourse analysis were explicit. Pedagogy/intervention: Online course activity analyzing Colombian and Chilean presidential speeches Technology: Zoom and online course materials	Political-speech transcripts, discourse-analysis categories, course activity artefacts/discussion	The analysis illustrates how political discourse can be used to connect scientific, political and ethical dimensions of climate-change education.	Some concerns. Use for critical-pedagogy and context; not evidence that critical thinking improved
Irawati & Sulisworo (2023) Journal article; Indonesia	Lower secondary education (Grade 7); Science: water quality and	Quasi-experimental pretest-posttest control-group	Emphasis: Critical thinking Coverage: Evidence analysis, problem-	PBL and IoT-supported authentic-data learning were explicit; construct alignment was	20-item prior-knowledge test, five critical-thinking essays, 25-	The PBL+IoT group outperformed regular PBL; learning interest contributed to	Minor concerns. Core evidence with non-randomized

Study and publication context	Educational context and sample	Study design and MMAT category	HOT construct emphasis and coverage	Framework, pedagogy, and technology	Assessment and analysis	Main synthesis finding	MMAT appraisal and appropriate synthesis use
	pollution 57 students from two schools/classes (32 PBL+IoT; 25 regular PBL)	design with covariates MMAT category: Quantitative non-randomized	solving, decision-making and collaborative/metacognitive processes	demonstrated through tasks and essays. Pedagogy/intervention: PBL with smart-water-monitoring IoT versus regular PBL Technology: IoT water-quality sensors/application	item learning-interest questionnaire, water-quality worksheet; covariance-based analysis	critical-thinking outcomes.	and cluster caveats
Zuhaida (2018) Conference proceeding; Indonesia	Higher education / preservice science teachers; Basic chemistry laboratory 55 students (29 metacognitive experiment; 26 conventional laboratory)	Quasi-experimental pretest-posttest control-group study with descriptive performance assessment MMAT category: Quantitative non-randomized	Emphasis: Problem-solving and metacognition Coverage: Declarative/procedural knowledge, planning, information management, monitoring, debugging, evaluation and argumentation	Schraw–Dennison metacognition and problem-based laboratory learning were explicit. Pedagogy/intervention: Metacognition-based basic-chemistry experiment versus standard laboratory procedure Technology: Laboratory experimentation	Problem-solving report/presentation rubrics, concept-description test, student responses; N-gain and descriptive analysis	The experimental group achieved high problem-solving rubric scores and greater concept gains than the comparison group.	Major concerns. Use cautiously; avoid precise causal claims

Note. All 24 eligible studies were retained. Methodological appraisal was used to calibrate interpretation rather than determine inclusion. No included study operationalized SOLO Taxonomy as its primary empirical framework; SOLO should therefore be presented as a theoretical comparison framework, not as a dominant corpus finding.

Abbreviations: HOT = higher-order thinking; HOTS = higher-order thinking skills; MMAT = Mixed Methods Appraisal Tool; PBL = problem-based learning; PjBL = project-based learning; SSI = socioscientific issues; IoT = Internet of Things; R&D = research and development.