

Systematic Review

Effective Strategies for Learning Solving Well-Structured Problems in STEM Education: A Systematic Literature Review

Gabriel Taban ^{1,2,*} , Jorrick Beckers ^{3,4} , Felicitas Biwer ², Kelly Beekman ³  and Anique B. H. de Bruin ² 

¹ Fontys Technology, Fontys University of Applied Sciences, 5612AP Eindhoven, The Netherlands

² Department of Educational Research & Development, School of Health Professions Education, Maastricht University, 6200MD Maastricht, The Netherlands

³ Department of Education, Fontys University of Applied Sciences, 5612MA Eindhoven, The Netherlands

⁴ Faculty of Educational Science, Department of Social Learning, Open University of the Netherlands, 6401DL Heerlen, The Netherlands

* Correspondence: g.taban@fontys.nl

Abstract

Problem-solving is central to STEM in higher education, yet evidence on effective learning strategies for problem-solving remains fragmented across cognitive psychology and discipline-specific research. This systematic literature review, conducted following the PRISMA framework, identifies, categorizes, and analyzes empirical evidence for the learning of well-structured problems in STEM higher education. The strategies reviewed build on two established frameworks of desirably difficult and generative learning strategies. Across three databases (ERIC, PsycINFO, Web of Science), supplemented by citation tracking and screening with ASReview of over 11,000 records, 34 peer-reviewed empirical studies were retained. This review addresses which strategies improve problem-solving in STEM higher education, and how student characteristics, teacher support, and implementation factors shape their effectiveness; data were extracted thematically across these four dimensions. Results show that distributed practice, drawing, enacting, self-explanation, and testing improved problem-solving across STEM contexts. The evidence for interleaving and concept mapping was less consistent. The size and reliability of effects depend on how each strategy was implemented. Effects were strongest when students actively generated content and when problem-solving was measured on delayed assessments. Particularly for testing, distributed practice, and interleaving, effects were also larger when interventions were sustained over time and structurally embedded in courses, though this pattern was not equally evident across all strategies. The review highlights the need for broader coverage of disciplines and student populations. Evidence for teacher's support role is not present in any included study.



Academic Editors: Amber Simpson and Jason Morpew

Received: 25 August 2026

Revised: 12 October 2026

Accepted: 14 October 2026

Published: 16 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: STEM education; problem-solving; higher education; learning strategies; generative learning; desirable difficulties

1. Introduction

Problem-solving strategies play a critical role in STEM education. Students engage in diverse problem-solving tasks, such as designing experiments to establish causal links between air pollution and diseases, selecting appropriate thermal management systems for electric vehicle batteries, or determining optimal methods for solving complex mathematical equations. To manage such complexity, students need to develop problem-solving skills progressively, working from well-structured problems toward increasingly open-ended

challenges. For example, in an educational experiment in physics, [Adams and Wieman \(2015\)](#) identified more than forty skills involved in solving a complex problem. In STEM higher education, even the simplest problems can be complex: students must understand important principles, model complex situations, and connect disparate concepts. Research has shown that novice problem solvers typically rely on surface-level cues and spend little time analyzing the problem, whereas expert problem solvers focus primarily on the underlying concepts and principles, and devote more effort to understanding the problem's structure ([Chi et al., 1981](#)). Decades of research in cognitive science and educational psychology have identified learning strategies that enhance retention and transfer of knowledge and skills ([Adesope et al., 2017](#); [Agarwal et al., 2021](#); [Bisra et al., 2018](#); [Brunmair & Richter, 2019](#); [Leutner & Biele, 2025](#); [Nesbit & Adesope, 2006](#)). The existing reviews either examine strategies in general educational contexts or focus on retention of factual knowledge, rather than specifically on the learning of problem-solving in STEM disciplines. This distinction matters: problem-solving requires not only retention but integration, transfer, and strategic application of knowledge under novel conditions. Whether these strategies translate effectively into improvements in STEM problem-solving, however, is a question that has not been systematically answered.

To understand how students can effectively navigate this complexity, we must first establish what constitutes a problem and problem-solving. Following the work of [Newell and Simon \(1972\)](#), [Gick \(1986\)](#), and [Jonassen \(2000\)](#), all anchored in the information processing theory ([Atkinson & Shiffrin, 1968](#); [Baddeley & Hitch, 1974](#)), in this review we define a problem as the difference between a current state and a goal state, and problem-solving as “any goal-directed sequence of cognitive operations” ([Anderson, 1980](#)). More specifically, this goal-directed process is most tractable when both the initial state and goal state are well-defined, the hallmark of well-structured problems. [Jonassen \(2000\)](#) distinguishes between well-structured and ill-structured problems. Well-structured problems, typically found at the end of textbook chapters, present all problem elements to students and require them to apply a limited number of concepts, rules, and principles, all organized in predictable ways. Ill-structured problems contain problem elements that are unknown and should be found from other external sources and contain multiple or even no solution at all.

Students face significant challenges when confronting problems of either type. Students struggle when trying to solve problems because complexity requires them to apply knowledge from different areas of their coursework, often spanning multiple semesters or even years ([Ibrahim et al., 2017](#); [Badeau et al., 2017](#)). Consequently, they often adopt flawed approaches to problem-solving. A common approach is briefly analyzing what has been given (current state) and what the result should be (goal state), then trying to reduce the gap by immediately attempting to apply formulas, sometimes called the ‘plug-and-chug’ approach ([Bergqvist, 2007](#); [Walsh et al., 2007](#)). Their ‘plug-and-chug’ approach contrasts sharply with effective problem-solving, which requires students to systematically analyze concepts, strategically plan solutions, methodically implement strategies, and reflectively evaluate results.

Developing effective problem-solving skills requires students to acquire both procedural knowledge (what actions to perform and how to perform them) and conceptual knowledge (why those actions solve the problem and when to apply them) ([Rittle-Johnson & Alibali, 1999](#)). Traditional approaches to teaching problem-solving have often equated practice volume with skill development: if students solve enough problems, they will learn to solve problems. This assumption underlies conventional instructional formats that emphasize supervised practice sessions and procedural heuristics (e.g., Polya's problem-solving steps ([Polya, 1945](#))). While practice helps students both automate pro-

cesses and strengthen skills, students who rely on procedures without conceptual understanding often struggle to transfer their knowledge when problem structures change (Rittle-Johnson & Alibali, 1999).

In recent decades, research in cognitive and educational psychology has provided robust evidence for cognitive learning strategies that effectively enhance long-term retention and transfer of knowledge. One concept related to learning strategies is that of ‘desirable difficulties’ (Bjork & Bjork, 2014; Soderstrom & Bjork, 2015), which are learning conditions that make initial learning more difficult, but result in superior long-term retention and greater chances of transfer of knowledge. While learners generally prefer approaches which make learning feel easier, such as cramming before the test or rereading texts, learning outcomes are higher when learning contains desirable difficulties. The ‘difficulty’ is desirable largely because it encourages students to process information effortfully, which strengthens learning. Complementing the concept of desirable difficulties, another concept is that of ‘generative learning strategies’ (Fiorella & Mayer, 2016). These strategies involve learning activities that require students to actively make sense of new information by selecting, organizing, and integrating it with their prior knowledge. Research suggests that teachers are more effective when they actively facilitate learning (Freeman et al., 2014; Lima et al., 2017). Moreover, students learn better when they actively engage in the learning process (Nguyen et al., 2021; Freeman et al., 2014; Theobald et al., 2020). In doing so, these strategies transform students from being passive recipients into active participants who must actively process information to understand it in a purposeful and coherent way.

Given the robustness of evidence for the effectiveness of desirably difficult and generative learning strategies on retention and transfer of knowledge and skills (Roediger & Pyc, 2012; Agarwal et al., 2021; Q. Zhang & Fiorella, 2021; Carpenter et al., 2022; Leutner & Biele, 2025) and their potential for application in STEM education (Samani & Pan, 2021; Ruitenburg et al., 2025; Rohrer & Taylor, 2006), we chose to focus on these two categories of learning strategies for the current review. Two studies in particular represent the dominant strands in scientific discussion on this theme. Dunlosky et al. (2013) reviewed ten learning techniques and rated their relative utility across student characteristics, materials, and criterion tasks, with practice testing and distributed practice receiving the highest ratings. Fiorella and Mayer (2016) introduced a complementary framework of eight generative learning strategies, grounded in Wittrock’s generative model of comprehension (Wittrock, 1974) and Mayer’s select—organize—integrate model (Mayer, 2014). Although desirable difficulties and generative learning derive from different theoretical traditions, we combine them deliberately, for three reasons. First, the two converge on a shared principle: learning is enhanced when students engage in effortful, active processing rather than passively receiving information. Second, the two sources overlap at the level of individual strategies (self-explanation, summarizing, imagining, and self-testing appear in both Dunlosky et al. (2013) and Fiorella and Mayer (2016)). A separation is neither possible nor useful. Third, most important here, problem-solving requires both the retrievable knowledge emphasized by desirable difficulties and the integration and transfer emphasized by generative learning. We treat the two frameworks as complementary.

Research on how interventions are implemented shows they can vary considerably depending on student characteristics, teacher behavior, and contextual factors (Agarwal et al., 2021; Leutner & Biele, 2025). These factors are crucial for understanding when and for whom these strategies work. First, student characteristics play a determining role in how they benefit from these strategies when learning problem-solving. Prior knowledge appears essential in this regard: whereas novices may become overloaded by too demanding challenges, advanced learners may benefit from more cognitively demanding tasks (expertise reversal effect (Kalyuga, 2007)). Additionally, whether students

feel motivated appears to influence if they persevere in the face of complex challenging problems or give up. Second, teacher support is crucial for successful implementation. Teachers must balance desirable difficulties carefully: too much assistance can lead to shallow processing and lack of engagement of long-term memory, while too little assistance leads to confusion, frustration, and wasted time (Koedinger & Alevan, 2007). When teachers instruct effectively, students are more likely to pursue appropriate learning strategies. Third, implementation factors play a role in practice. Instructors must carefully calibrate task complexity to students' level to prevent cognitive overload. Whether students can access materials or technology also influences whether these strategies can realistically be implemented in educational practice.

Given these multiple influencing factors, it is crucial to examine the empirical evidence for strategy effectiveness in problem-solving research. The effectiveness of these 'desirable difficulties' and 'generative learning strategies' on learning and performance has been demonstrated in numerous systematic reviews and meta-analyses in different contexts and educational levels. In problem-solving research specifically, three obstacles, however, complicate the picture: fragmentation, limited transfer validity, and a persistent gap between research and practice. First, evidence on these strategies is scattered across either cognitive psychology or the specific STEM-education literature. As a result, converging evidence remains invisible to those working at the congruence of these fields and limits them in accumulating usable knowledge. Second, teachers and curriculum designers invoke 'effective learning strategies' as evidence-based practice, yet much of this evidence was built on non-STEM outcomes like retention of knowledge or vocabulary learning. Problem-solving is, however, more complex. Without synthesizing the evidence specifically for problem-solving outcomes in STEM-education, recommendations risk a validity gap: strategies effective for simple retention tasks may not transfer to the more demanding requirements of STEM problem-solving. Third, partly because the available evidence has not been synthesized into a form teachers can directly use into practice, STEM instruction continues to rely on 'plug-and-chug' routines. Therefore, a dedicated systematic review of the merits of the various strategies in terms of learning problem-solving in STEM education is needed.

Our systematic review seeks to bridge this gap by identifying, categorizing, and analyzing evidence-informed strategies that demonstrably improve learning problem-solving across various STEM contexts, and addresses the following research question:

What desirably difficult and generative learning strategies improve the learning of solving well-structured problems in STEM education and to what extent is this influenced by students' characteristics, teachers' support, and other implementation factors?

The Present Study

This review aims to accomplish four objectives: (1) identify and categorize evidence-informed strategies that enhance problem-solving learning in STEM disciplines, (2) analyze the effectiveness of different strategies across various STEM contexts and student populations, (3) synthesize findings regarding optimal conditions for implementing effective problem-solving instruction, and (4) identify gaps in current research that justify future investigation.

We focus specifically on well-structured problem-solving within STEM education settings at the level of higher education. We exclusively examine strategies that have been empirically tested. We synthesize research from cognitive psychology, educational psychology, and discipline-specific education research to provide a comprehensive understanding of effective approaches to learning problem-solving.

In this review, the term 'effective learning strategies' will refer collectively to both 'desirable difficulties' and 'generative learning strategies': distributed practice, drawing,

elaborative interrogation, enacting, highlighting, imagining, interleaved practice, keyword mnemonic, mapping, rereading, self-explaining, self-testing, summarizing, and learning by teaching. A specific definition of each strategy is provided in the results section. This categorization follows two landmark studies that provide both theoretical grounding and empirical validation for these strategies: Dunlosky et al. (2013), who rated strategy utility across learning conditions next to determining strategy effectiveness for long-term learning and transfer, and Fiorella and Mayer (2016), whose generative learning framework integrates the cognitive mechanisms explaining strategy effectiveness.

Building on this foundation, this systematic review makes three important contributions. First, this review serves as a resource for bridging the gap between scientific knowledge about learning and actual classroom practice in STEM education. Second, it will establish a foundation for future research by identifying gaps in current knowledge and promising directions for investigation. Third, it will provide students and teachers with evidence-based guidance for selecting and implementing effective problem-solving strategies.

2. Methods

Both the review process and the reporting of results followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021). The review protocol was pre-registered on the Open Science Framework (OSF) prior to data extraction [<https://osf.io/hxscd>, accessed on 13 October 2026]. Data sets can be retrieved via [<https://osf.io/zh3vr>, accessed on 13 October 2026]. The review process began with the formulation of a research question to guide the study. Inclusion and exclusion criteria were then specified to define the search strategy. Relevant studies were identified across three major databases and subsequently screened according to the established selection criteria. Finally, a thematic synthesis was conducted on included full-text articles to categorize the effects of different learning strategies on problem-solving outcomes and to describe their key implementation features, student characteristics, and instructional conditions. The details of the procedure in each step are given in the next sections.

2.1. Selection Criteria

Based on the research aim and the proposed research questions, we adopted the inclusion and exclusion criteria detailed in Table 1.

Table 1. Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
Participants at higher education level	No studies published as reviews, meta-analyses, conference proceedings, opinions, reports, books, or non-peer-reviewed literature
Participants from STEM education programs	No qualitative studies, case studies, or observational studies
Studies in STEM education	No studies on solving ill-structured problems
Studies on solving well-structured problems	
Studies with a control group in the design	
Effect on learning outcomes when solving problems in a STEM-field as dependent variable	

Table 1. Cont.

Inclusion Criteria	Exclusion Criteria
At least one of the following strategies as independent variable: distributed practice, drawing, elaborative interrogation, enacting, highlighting, imagining, interleaved practice, keyword mnemonic, mapping, rereading, self-explaining, self-testing, summarizing, teaching	
Studies published in English	
Only studies published in peer-reviewed journals	

2.2. Literature Search
2.2.1. Identification of Studies via Databases

To cover the field of educational and psychological research comprehensively, three relevant literature databases were selected and searched systematically for eligible studies: ERIC, PsycINFO, and Web of Science. The search strategy was designed around four key topics that aligned with our research question: problem-solving, STEM, learning strategies, and higher education. A unique search string was defined for each database. The strings are presented in Appendix A.

After removing duplicates and ineligible document types, a total of 11,085 studies remained for screening, the search and identification process was initially completed in October 2024. A follow-up search was conducted six months later in April 2025.

The process of selecting literature is visualized in the PRISMA flowchart (see Figure 1).

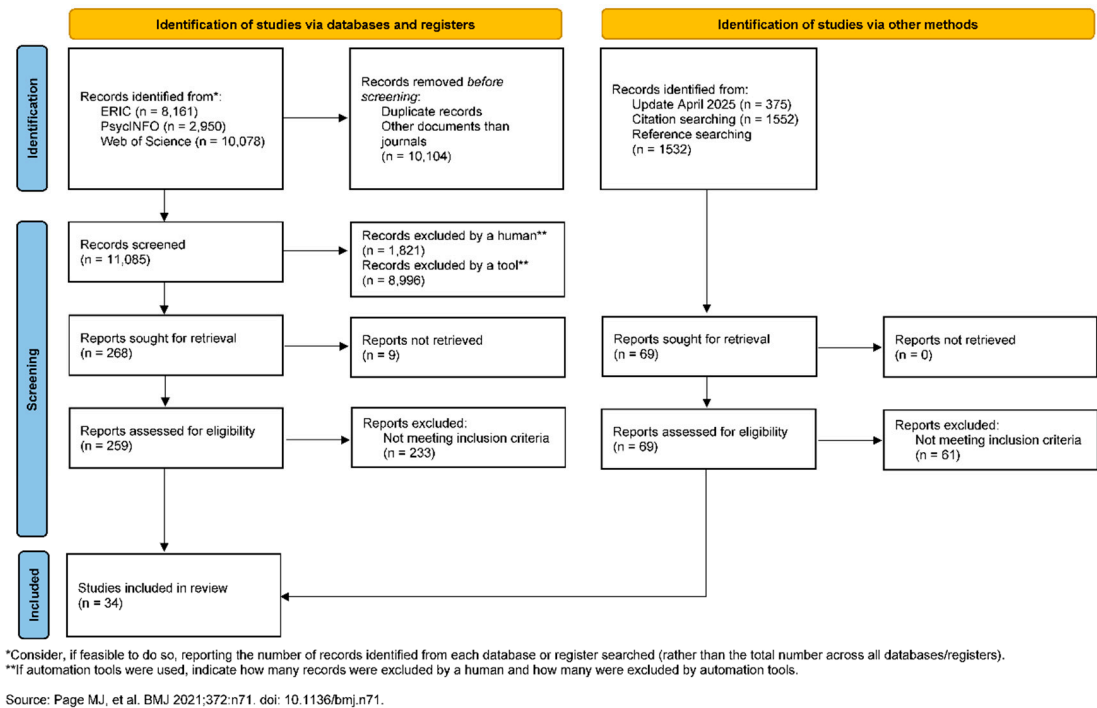


Figure 1. The process of selecting articles follows the PRISMA framework (Page et al., 2021).

2.2.2. Screening Procedure

Given the large number of records retrieved, ASReview version 1.6.3 was used to support the screening process (Van de Schoot et al., 2021). ASReview is an open-source tool developed by the ASReview Innovation Lab at Utrecht University designed to improve the efficiency of systematic literature screening. It applies active learning and multiple ma-

chine learning models within a Human-in-the-Loop approach (Van de Schoot et al., 2021). Researchers make inclusion and exclusion decisions while the algorithm continuously refines its predictions. Starting with initial categorizations, ASReview sequentially presents records, learning from each researcher's decision to prioritize the most relevant remaining records.

Especially when working with large datasets, a critical challenge in active learning is determining the optimal stopping point. On one side, if a researcher screens the database excessively, the advantage of an active learning tool disappears: it takes too long and provides only a minimum number of additional relevant records. On the other side, stopping too early risks missing relevant records (Yu et al., 2018). The SAFE procedure (Boetje & van de Schoot, 2024) addresses this challenge by integrating multiple heuristics into a practical stopping rule. Its four-phase approach balances efficiency with reliability, minimizing the risk of overlooking key studies and ensuring thoroughness: Phase 1: Screen a random training set; Phase 2: Apply active learning; Phase 3: Find additional relevant records using alternative models; and Phase 4: Evaluate screening quality.

Phase 1: Screen a random training set. The machine learning model required training data, meaning records labeled as relevant or irrelevant for its initial training. To create this dataset, we selected a representative sample of 150 studies (slightly exceeding the proposed 1% of the total records), comprising the 50 oldest, 50 newest, and 50 randomly selected records from the full dataset of 11,085 articles. In addition, six pre-identified key studies—studies already confirmed as relevant—were included to guide the model. Titles and abstracts were screened independently by three reviewers, yielding 11 relevant records in the initial sample. Based on this initial screening, a rough estimate of 811 total relevant records (RR_T) was established, which served as a reference for the stopping heuristic in Phase 2.

Phase 2: Active learning. Using the labeled data from Phase 1, the model prioritized studies most likely to be relevant. Feature extraction, the transformation of text data into a numerical format, was performed with the term-frequency-inverse document frequency (TF-IDF) algorithm (Salton & Buckley, 1988). The classification, sorting data into different categories, was conducted using the Naive Bayes Classifier algorithm (Gomes et al., 2017; Jackson & Moulinier, 2007). Screening continued until all four predetermined stopping criteria were satisfied: (1) All key studies had been identified as relevant; (2) At least twice the RR_T records had been screened; (3) A minimum of 10% of the total dataset had been screened; (4) No relevant records had been identified in the most recent 75 records. After screening 1700 records, all four criteria were satisfied, yielding 237 records classified as relevant. These results were subsequently discussed and validated by three reviewers.

Phase 3: Deep learning. To identify potentially missed studies due to conceptual ambiguity in abstract texts, a more advanced model was employed. To identify possible relevant studies, the algorithm must “dig deeper” into a text to find its essence. Feature extraction was performed with the bidirectional encoder representations from transformers (sBERT) algorithm (Reimers & Gurevych, 2019) and the classification was conducted with the Random Forest algorithm (Breiman, 2001). Since deep learning models require more training data, the algorithm was retrained on the labeled data from Phase 2 and reordered the remaining unscreened records using deeper semantic analysis. Screening stopped after a cumulative total of 2087 records (including Phase 2), when no new relevant studies appeared in the most recent 75 screened. This process yielded 29 additional records in a revised total of 266. The results were then discussed and validated by three reviewers.

Phase 4: Evaluation of quality. To minimize the risk of erroneously excluding relevant studies, previously discarded records were re-screened using the initial model (TF-IDF + Naive Bayes), the model being re-trained on 10 relevant and 10 irrelevant records

from earlier phases. A second reviewer assessed the most likely to be false negatives and screening was stopped when no new relevant studies had been identified in the most recent 75. This phase yielded 2 additional relevant studies, resulting in a final total of 268.

Full-text analysis. The 268 selected studies were analyzed for relevance according to the predefined selection criteria. The first author read all studies, with two other reviewers giving second opinions where necessary. Following this process, 26 studies were labelled as relevant.

2.2.3. Identification of Sources via Other Methods

The updated search done in April 2025 identified an additional 375 records. Additionally, some relevant studies may not be captured only through the mentioned search terms and criteria. Citation tracking was therefore used to identify additional studies within the same themes. Citation tracking was performed using Citationchaser ([Haddaway et al., 2022](#)), with the 26 records deemed relevant after full-text analysis serving as the starting point. A total of 1532 records were found as references and 1552 records were found as citations. All newly identified records (update, references, and citations) were combined into a single data set and screened with ASReview. Feature extraction was performed with the sBERT algorithm ([Reimers & Gurevych, 2019](#)), and the classification was conducted with the Random Forest algorithm ([Breiman, 2001](#)). The model was trained on the labeled data from Phase 3. Screening proceeded until the predefined stopping criteria were met, which occurred after 369 records. This phase identified 69 additional studies. After the full-text analysis, an additional number of 8 studies were added for the final review. The results were discussed by three reviewers. A total of 34 studies were included in this review (26 via the original search and 8 added via additional searches).

Data was extracted by the first author. A second reviewer verified the extracted data for a random subset of studies against the source articles. Disagreements were resolved by discussion until consensus was reached, where needed in collaboration with the rest of the research team. For each experiment, we extracted bibliographic details, STEM domain, educational level, sample size, experimental design, learning strategy investigated, control condition, problem-solving outcome measure(s), key findings, effect sizes, and implementation characteristics. We also added a short qualitative description of the eligibility of each study. Table 2 presents an overview of the evidence per learning strategy. In Tables S1–S7 of the Supplementary Materials, we included all these experimental details.

Table 2. Overview of evidence per learning strategy.

Strategy	Nr. Studies (nr. Experiments)	Risk-of-Bias (≤ 6 /nr. Studies)	STEM Domain (nr. Studies)	Study Design (nr. Experiments)	Outcome Measured	Intervention & Key Implementation Conditions	Comparison Conditions(s)	Effect Direction (nr. Experiments)
Distributed practice	4 (4)	0/4	Mathematics 4 (Hopkins et al., 2016; Lyle et al., 2020, 2022; Y. Zhang et al., 2024)	W 3 R 1	(Hopkins et al., 2016) (PS) (Lyle et al., 2020) (PS) (Lyle et al., 2022) (PS) (Y. Zhang et al., 2024) (PS)	Quiz questions distributed across weeks; worked examples split over two sessions one week apart. Inter-practice interval of at least one week; practice redistributed with existing course. Benefits on delayed posttest.	Massed practice 4	4 + /0 → /0 –
Drawing	10 (12)	1/10	Physics 6 (Bahtaji, 2020; Heckler, 2010; Kohl et al., 2007; Maries & Singh, 2018a, 2018b, 2023); Chemistry 2 (Padalkar & Hegarty, 2015; Stieff et al., 2016b) Biology 1 (Heideman et al., 2017) Engineering 1 (Wu et al., 2020)	Q 7 R 2 W 1 C 2	(Bahtaji, 2020) (KC) (Heckler, 2010) (PS) (Heideman et al., 2017) (KC) (Kohl et al., 2007) (PS) (Maries & Singh, 2018a) (PS) (Maries & Singh, 2018b) (PS) (Maries & Singh, 2023) (PS) (Padalkar & Hegarty, 2015) (KC) (Stieff et al., 2016b) (KC) (Wu et al., 2020) (KC)	Students construct a diagram, graph or representation while solving problems, either prompted during solving or embedded in an instructional sequence. Students generate the representation themselves, providing diagrams performed worse than no diagram; the benefit is confined to difficult problems; drawing quality predicts performance more than presence of drawing.	No prompt, no support, or instructor provided diagram 5; ready-made graphs 1; alternative scaffold 4; standard instruction 2	7 + /3 → /2 –
Enacting	3 (4)	0/3	Chemistry 2 (Ping et al., 2022; Stieff et al., 2016a) Engineering 1 (Kadam et al., 2021)	R 4	(Kadam et al., 2021) (PS) (Ping et al., 2022) (PS) (Stieff et al., 2016a) (PS)	Students produce the movement rather than observe it; perform tasks themselves through digital manipulation of 3D objects; physical model manipulation; mental imagery. Scaffolded sequence demonstration—joint performance— independent practice. Gesture training transfers to assessment without models present.	Observed gesture/text script 1; concrete model instruction 1; conventional instruction 1; three active modalities compared 1	2 + /2 → /0 –

Table 2. Cont.

Strategy	Nr. Studies (nr. Experiments)	Risk-of-Bias (≤ 6 /nr. Studies)	STEM Domain (nr. Studies)	Study Design (nr. Experiments)	Outcome Measured	Intervention & Key Implementation Conditions	Comparison Conditions(s)	Effect Direction (nr. Experiments)
Interleaving	3 (3)	1/3	Physics 1 (Samani & Pan, 2021) Mathematics 1 (Gorgievski & DeFranco, 2024) Chemistry 1 (Gulacar et al., 2022)	R 2 W 1	(Gorgievski & DeFranco, 2024) (PS) (Gulacar et al., 2022) (PS) (Samani & Pan, 2021) (PS*)	Problem sets in which problems from different topics alternate, delivered as homework assignments or as problem-solving sessions. Intervention is sustained and frequent for weeks; mixing spans the whole course, not just a single chapter; cramming before scheduled exams masks the effect, detection requires an unannounced delayed test.	Blocked practice with the same problems 3	1 + /2 → /0 –
Mapping	1 (1)	1/1	Biology 1 (Okebukola, 1992)	Q 1	(Okebukola, 1992) (PS*)	After one week of training, students construct concept maps during every biology lesson for six months, listing concepts, arranging them hierarchically from abstract to concrete and labelling the connecting links, with summary maps every three weeks. Benefit in the mapping act itself.	Traditional instruction without mapping 1	1 + /0 → /0 –
Self-explanation	5 (7)	0/5	Physics 3 (Badeau et al., 2017; Gjerde et al., 2022; Hausmann & VanLehn, 2010) Chemistry 1 (Ionas et al., 2012) Mathematics 1 (Miller-Cotto & Auxter, 2021)	R 6 Q 1	(Badeau et al., 2017) (PS*) (Gjerde et al., 2022) (PS) (Hausmann & VanLehn, 2010) (PS*) (Ionas et al., 2012) (PS) (Miller-Cotto & Auxter, 2021) (KC)	Students explain worked examples or own reasoning in own words, in writing, verbally after each segment video-based worked examples, or by answering guiding questions before attempting a problem. Prompts direct students to underlying principles and conditions of application; open ‘why’ prompts harm performance; worked examples with deliberately omitted justifications; below prior-knowledge threshold strategy hinders rather than helps.	No training 2; problem-solving only 1; paraphrasing 1; unguided problems 1; business-as-usual problem-solving 1; SE without prior retrieval 1	5 + /1 conditional (=positive only above a prior-knowledge threshold)/1 –

Table 2. Cont.

Strategy	Nr. Studies (nr. Experiments)	Risk-of-Bias (≤ 6 /nr. Studies)	STEM Domain (nr. Studies)	Study Design (nr. Experiments)	Outcome Measured	Intervention & Key Implementation Conditions	Comparison Conditions(s)	Effect Direction (nr. Experiments)
Testing	8 (9)	2/8	Physics 3 (Gjerde et al., 2020; M. Zhang et al., 2023; Zu et al., 2019) Mathematics 3 (May, 2024; Muzsnay et al., 2024; Szabó et al., 2023) Biology 1 (Bailey et al., 2017) Engineering 1 (Morphew et al., 2020)	C 3 R 2 Q 2 W 2	(Bailey et al., 2017) (PS) (Gjerde et al., 2020) (PS + KC) (May, 2024) (PS + KC) (Morphew et al., 2020) (PS + KC) (Muzsnay et al., 2024) (PS) (Szabó et al., 2023) (PS) (M. Zhang et al., 2023) (PS) (Zu et al., 2019) (PS*)	Structurally embedded tests and retrieval with immediate or delayed feedback; delayed second-chance retake. Students retrieve studied material instead of restudy; frequent low- or moderate-stakes tests across semester, as a single retrieval session, as retrieval distributed across lectures, or as online practice exams; immediate item-level feedback combined with a delayed second-chance retake.	Pretest, within group 1; traditional exam 3; prior cohort without intervention 1; alternative practice-exam format 1; study of worked problems 1; peer discussion restudy 1; worked examples 1.	8 +/1 $\rightarrow$ 0 –

Notes. Number of studies (number of experiments): several studies contributed more than one experiment (34 studies, 40 experiments). Risk-of-bias is at study level (total 34) and gives information about the number of studies assessed with 5 or 6 points. Study design: R = randomized between-subjects; Q = quasi-experimental; W = within-subjects; C = between-cohort or between-semester comparison. Outcome measured: PS* = problem-solving on novel/transfer problems; PS = problem-solving on practiced problems; KC = knowledge or course-level outcome. Effect direction, coded per experiment: + statistically significant advantage for the strategy; $\rightarrow$ no statistically significant difference; – statistically significant disadvantage for the strategy. Full details at experiment level are given in Supplemental Materials Tables S1–S7.

3. Results

A total of 34 studies were included in this review. Results are presented thematically by strategy in alphabetical order. For each strategy, we describe the overall pattern of effects on problem-solving performance, followed by implementation factors, and the role of student characteristics and instructional support where data permit.

For each study, Table 2 provides the type of outcome that was measured. We distinguished between two types of problem-solving performance outcomes (novel/transfer problems or practiced problem types), and outcomes that measured components or correlates of problem-solving, such as factual or conceptual knowledge, representational competence, and aggregate course or examination scores. Our strategy-level conclusions therefore rest on problem-solving outcomes, and most strongly on novel and transfer tasks, with knowledge- and course-level results reported as convergent rather than primary evidence.

3.1. Risk-of-Bias Appraisal

The included studies vary in design and outcome measures. Therefore, we conducted a risk-of-bias appraisal using a ten-item checklist adapted from three instruments developed for systematic reviews in health-care and health-professions education (Downs & Black, 1998; Buckley et al., 2009; Harrison et al., 2021). Each study was scored 0 or 1 on ten quality indicators: research question, theoretical grounding, participants, statistical power, data-collection method, analysis, reporting of results, reproducibility, discussion of strengths and limitations, and justification of conclusions. The full appraisal is presented in Appendix B. These results are integrated into Table 2. Scores ranged from 5 to 10 ($M = 8.6$), with 29 of 34 studies (85%) scoring 7 or higher. One study scored 5 points (Bahtaji, 2020) and four studies scored 6 points (Bailey et al., 2017; Gulacar et al., 2022; May, 2024; Okebukola, 1992). Following (Buckley et al., 2009), higher quality studies were considered those which met a minimum of 7 points. Findings from the five lower-scoring studies are reported with caution in the sections that follow.

3.2. Distributed Practice

Distributed practice or spacing is considered a desirable difficulty and involves distributing learning over time across different study sessions (Bjork & Bjork, 2014). Dunlosky et al. (2013) rated distributed practice as one of only two learning techniques (alongside practice testing) at highest utility.

Across the four studies reviewed (Hopkins et al., 2016; Lyle et al., 2020, 2022; Y. Zhang et al., 2024), distributed practice appears to consistently improve students' performance on mathematics problem-solving tasks, with effects ranging from small-to-medium (Lyle et al., 2022; Y. Zhang et al., 2024) to medium-to-large (Hopkins et al., 2016; Lyle et al., 2020). All four studies reported statistically significant benefits of distributed over massed practice on retention tests, which provides preliminary converging evidence from both classroom and experimental settings, although the small number of studies and the sole focus on mathematics content limit the generalizability of this conclusion.

Three studies were conducted in authentic engineering undergraduate classrooms, providing ecological validity. Hopkins et al. (2016) and Lyle et al. (2020) focused on a precalculus course, while Lyle et al. (2022) extended the investigation to a calculus course. In all three, distributed practice was implemented through online quizzes in which questions for a given learning objective were either concentrated on a single quiz (massed) or distributed across multiple quizzes over several weeks. In contrast, Y. Zhang et al. (2024) distributed the learning phase itself during study of worked examples across two sessions separated by one week and assessed retention in a controlled experimental setting.

An important finding is that the benefits of distributed practice on problem-solving tended to grow over time. [Y. Zhang et al. \(2024\)](#) found no advantage for the distributed group on an immediate posttest, with the benefit emerging only one week later, when massed learners had forgotten considerably more than the distributed learners. A similar pattern emerged in the classroom studies, where the advantage of distributed practice was clear in delayed assessments at end-of-semester exams. Importantly, [Hopkins et al. \(2016\)](#) demonstrated that these benefits were not limited to a single course: students who had experienced distributed practice in precalculus continued to outperform their massed peers in the following calculus course. This cross-semester persistence was replicated by [Lyle et al. \(2020\)](#), who further showed that only distributed practice produced durable gains.

The reviewed studies point to several implementation factors that appear to be associated with whether distributed practice benefits learning: timing and length of inter-practice interval, distinction between practice phase and learning phase, and short-term costs during practice itself.

With respect to the timing and length of distributed practice, the four studies are consistent in suggesting that a sufficient time interval between practice opportunities may be necessary for the effect to emerge, although the optimal interval remains unclear from this limited set of studies. In the three classroom studies, questions for a given learning objective were distributed across quizzes with intervals of one to three weeks, embedded within the regular course. This schedule required no additional study time from students, and the performance gains were achieved by redistributing when practice occurred, not by increasing the amount of practice. [Lyle et al. \(2020\)](#) tested this difference by independently manipulating both amount and spacing of practice. Their results showed that while increasing the amount of practice also helped within-semester retention, only distributing practice produced gains that persisted into the following semester.

In three of the four studies, distributed practice was applied to the practice phase: students were first introduced to the content in class, and questions on that content were subsequently distributed across multiple sessions ([Hopkins et al., 2016](#); [Lyle et al., 2020, 2022](#)). Taking a different approach, [Y. Zhang et al. \(2024\)](#) distributed the learning phase, splitting students' initial introduction to worked examples across two sessions separated by a week. Despite this difference, benefits of distributed practice were observed in all four studies, tentatively suggesting that distributing learning over time may extend beyond the practice phase to the initial learning phase as well.

With respect to the short-term costs that distributed practice introduces during practice, the distributed group performed worse on individual quiz questions than the massed group, even though their end-of-semester retention was superior ([Hopkins et al., 2016](#); [Lyle et al., 2020, 2022](#)). [Lyle et al. \(2022\)](#) extended this finding and showed that the difficulty was concentrated on the first questions and diminished progressively across subsequent questions. This pattern is consistent with the interpretation that distributed practice delays the consolidation of learning. A comparable pattern was observed by [Y. Zhang et al. \(2024\)](#), whose distributed group performed no better than the massed group immediately after learning. Taken together, these findings suggest that short-term performance can be misleading during distributed practice, and that delayed assessment appears to be required to reveal the full benefits.

The four studies offer very limited and possible noteworthy information regarding the role of student characteristics, specifically gender and prior knowledge. With respect to gender, only [Hopkins et al. \(2016\)](#) explicitly tested it as a moderator, finding a significant interaction that may suggest that distributed practice benefited male but not female students in the within-subjects analysis. However, the female subsample was very small. [Y. Zhang](#)

et al. (2024) enrolled an exclusively female sample, making gender-comparative conclusions impossible from that study. Regarding prior knowledge, only Y. Zhang et al. (2024) explicitly examined it directly, finding that prior knowledge predicted immediate but not delayed retention, and did not moderate the distributed practice effect itself. Given that prior knowledge was not examined in the remaining three studies, no general conclusions about its moderating role can be drawn at this stage.

3.3. Drawing

Drawing is a generative learning strategy in which learners produce by hand or with digital tools a pictorial representation that depicts the content of a situation (Fiorella & Mayer, 2016).

Across twelve experiments implementing drawing in varied contexts and with different instructional goals, the strongest problem-solving benefits emerged when students generated diagrams themselves rather than received ready-made ones. The most common implementation was to prompt students to draw a diagram during problem-solving and to compare their performance against students who received no such prompt (Heckler, 2010; Maries & Singh, 2018a, 2018b, 2023; Wu et al., 2020). Maries and Singh (2018a, 2018b, 2023) added a third condition in which students received a ready-made (“unproductive”) diagram. Other studies embedded drawing in larger instructional packages: concrete-model workshops (Stieff et al., 2016b), models-as-feedback training (Padalkar & Hegarty, 2015), sketching interventions (Heideman et al., 2017), drawing embedded in a representationally rich curriculum (Kohl et al., 2007), and varied scaffolding of graph construction (Bahtaji, 2020).

The studies varied widely in setting, duration, and delay before outcome testing. Most experiments took place in authentic classroom settings during regular coursework and assessed students on actual quiz or exam items; the exceptions were Padalkar and Hegarty (2015, Experiments 1 and 2), conducted in the laboratory. Intervention duration ranged from a single session (Heckler, 2010; Maries & Singh, 2018a, 2018b; Padalkar & Hegarty, 2015) to multi-week classroom implementations (Kohl et al., 2007; Stieff et al., 2016b; Wu et al., 2020). Delays varied similarly: most studies relied on same-session performance. Whereas Maries and Singh (2018b, 2023) also assessed a delayed midterm weeks later without further intervention, Stieff et al. (2016b) tracked three-unit exams and a cumulative final at week 16, and Heideman et al. (2017) assessed retention over three days to five weeks.

Despite this diversity, the evidence converges on a clear pattern: student-generated diagrams support problem-solving, whereas provided diagrams often do not. In the three studies by Maries and Singh (2018a, 2018b, 2023), prompted students (PO) consistently outperformed students given a ready-made diagram (DO) by medium-to-large margins, and the DO group consistently performed worse than students given no support. These benefits extended to a delayed midterm weeks later where no drawing intervention was applied (Maries & Singh, 2018b, 2023), indicating that the effects persist beyond the intervention session itself. In laboratory settings, drawing combined with active feedback yielded the largest effects, though these were measured on representational competence (Padalkar & Hegarty, 2015). Classroom-based studies yielded smaller but still positive effects: Stieff et al. (2016b) found medium-to-large gains that did not carry over to course assessments, Wu et al. (2020) reported a selective effect, with drawing prompts improving conceptual-question performance but not procedural performance and Kohl et al. (2007) observed a robust association between complete, correct diagrams and performance on a challenging problem.

The reviewed studies point to several implementation factors that appear to be associated with whether drawing benefits learning: setting in which drawing is practiced

and the associated instructional scaffolding, quality of the drawings produced, difficulty of problems, and type of feedback students receive.

Classroom studies produced smaller but more ecologically meaningful effects than laboratory studies. What mattered for scaffolding was not how students were directed to draw, but whether they could check their drawings with a correct model. Most experiments took place in authentic classroom settings during regular coursework. The two exceptions, conducted in the laboratory, yielded the largest effect sizes, although on representational competence (Padalkar & Hegarty, 2015, Experiment 1 and 2). The classroom studies, however, measured retention on delayed midterms (Maries & Singh, 2018b, 2023) or semester-long examinations (Stieff et al., 2016b). Studies scaffolded the drawing activity to very different degrees: Bahtaji (2020) delivered explicit instruction in graph construction, Kohl et al. (2007) compared a strongly directed approach with a weakly directed approach, and Stieff et al. (2016b) scaffolded drawing through a two-week TA-led workshop series. Strongly and weakly directed approaches yielded comparable levels of representation use (Kohl et al., 2007), whereas scaffolding that actively engaged students in comparing their drawings with a correct model produced stronger outcomes (Stieff et al., 2016b).

Drawing quality predicted performance better than the simple act of drawing. Researchers often rated students' drawings with rubrics, and these studies found that the mere presence of a drawing was a weak predictor of performance; what mattered was whether the drawing captured task-relevant information. Maries and Singh (2018a, 2018b, 2023) classified diagrams as "productive" or "unproductive" according to whether they encoded directional or structural information relevant to the solution, while Padalkar and Hegarty (2015) used a three-level error typology (spatial, connectivity, fundamental). Kohl et al. (2007) rated free-body diagrams for presence, correctness, and labeling, and Stieff et al. (2016b) analyzed errors by representation type.

Drawings helped most with difficult problems. Kohl et al. (2007) reported that correct diagrams predicted success most strongly on a challenge problem designed to be very difficult to solve without a diagram, whereas trends on simpler recitation problems were much weaker. Maries and Singh (2018b) observed a parallel pattern, with effects strongest on a two-dimensional problem and negligible on a simpler one-dimensional variant.

Feedback on the drawing process had larger effects when students generated it themselves than when a tutor provided it. In Padalkar and Hegarty (2015, Experiment 2), students who received evaluative feedback from a tutor showed a strong pretest-posttest gain, yet students who generated their own feedback by aligning concrete models with their own drawings significantly outperformed the feedback-only group at posttest. Stieff et al. (2016b) built TA-mediated feedback into their workshops and reported medium-to-large gains on a representation-translation test that did not carry over to course exams. Wu et al. (2020) provided feedback from multiple sources (software, instructor, TA, peer coaches) and observed small gains on conceptual questions alongside increases in students' use of drawing.

Only one of the experiments addressed prior knowledge systematically. Maries and Singh (2023) compared introductory undergraduates with experienced graduate students on the same wave problem. Graduates outperformed undergraduates overall, and particularly on items requiring a diagrammatic approach, consistent with the view that expertise supports productive diagram use. Heckler (2010) contrasted honors and traditional students, finding that honors performed near ceiling regardless of condition, while traditional students were disadvantaged by being prompted to draw.

3.4. Enacting

Enacting involves engaging in task-relevant movements during learning, such as performing gestures or manipulating concrete objects. This happens in coordination with the lesson content (Fiorella & Mayer, 2016).

Despite variation in topic and intervention, spanning engineering drawing (Kadam et al., 2021) and organic chemistry (Ping et al., 2022; Stieff et al., 2016a), the three reviewed studies point to a clear trend: active generation of a spatial transformation consistently improves problem-solving performance, while passive exposure to the same transformation does not. Enacting was operationalized as digital manipulation of 3D objects to practice mental rotation for engineering drawing (Kadam et al., 2021), as physical model manipulation, hand gesture, or mental imagery applied to molecular rotation problems (Ping et al., 2022), and as gestures representing molecular spatial transformations (Stieff et al., 2016a).

The evidence for this is cumulative across studies. Effect sizes grew from negligible at the first posttest to large by the final posttest (Kadam et al., 2021), suggesting that enacting benefits may be sensitive to cumulative practice rather than immediate exposure. When the mode of active generation varied, all three conditions (imagine, action, and gesture) produced equivalent gains (Ping et al., 2022), indicating that once active generation is ensured, further differentiation by modality offers negligible additional benefit. Most directly, student-produced gestures outperformed both text instruction and observed gestures, with no significant difference between the latter two (Stieff et al., 2016a, Study 1), consistent with the interpretation that the critical variable is production, not exposure.

Three additional findings characterize the conditions under which these benefits are most robust. Greater spontaneous gesture use during independent problem-solving, observed among students who had received gesture-based instruction, was associated with higher performance (Stieff et al., 2016a, Study 1), suggesting that effective enacting instruction may alter how students approach spatial problems more broadly. Context independence of learning further distinguished gesture-trained from model-trained students (Stieff et al., 2016a, Study 2): the former maintained performance regardless of model availability at assessment, while the latter performed substantially worse without them. Finally, training completeness before assessment emerged as a boundary condition (Kadam et al., 2021): significant differences appeared only at tests following completion of the full training sequence, not at those administered after partial delivery. These conclusions should be interpreted cautiously, however, as they rest on three studies conducted in two disciplinary contexts.

The reviewed studies point to several implementation factors that appear to be associated with whether enacting benefits learning: timing and instructional structure.

Enacting proved effective in both single-session (Ping et al., 2022; Stieff et al., 2016a) and distributed multi-session formats (Kadam et al., 2021), indicating that duration matters less than how time is used. A single focused hour produced meaningful within-session gains, while effects accumulated progressively across posttests in the distributed design, suggesting that repeated enacting drives increasingly larger benefits.

A scaffolded sequence, from instructor demonstration through joint performance to independent practice, is a consistent feature of all three designs, suggesting that this progression may be a contributing factor to their effectiveness. Each begins with explicit modelling of the target spatial operation before students enact it themselves. That demonstration alone is insufficient is established most clearly by Stieff et al. (2016a, Study 1): participants who observed the instructor's gestures without producing them performed no better than those who read a text script. Demonstration without production did not make the strategy available for subsequent independent problem-solving.

Gender was explicitly examined only by [Ping et al. \(2022\)](#), where male and female students performed equivalently at both pretest and posttest, with no significant interaction between gender and either condition type or learning gain. As this finding derives from a single study, however, it should be treated as preliminary.

3.5. Interleaving

Interleaving is considered a desirable difficulty that involves alternating practice between different problem types within a single study session rather than completing all problems of one type before moving to the next type ([Dunlosky et al., 2013](#)).

Interleaving was tested as a strategy to improve problem-solving in physics ([Samani & Pan, 2021](#)), calculus ([Gorgievski & DeFranco, 2024](#)), and general chemistry ([Gulacar et al., 2022](#)), with different outcomes across the three studies. All three interventions contrasted interleaved problem sets with blocked sets in which problems from a single topic appeared consecutively. The effects varied between the studies: [Samani and Pan \(2021\)](#) found medium-to-large benefits, [Gorgievski and DeFranco \(2024\)](#) found no significant difference between blocked and mixed homework, and [Gulacar et al. \(2022\)](#) found no significant overall effect. Given that this review is based on only three studies, the small number limits the generalizability.

Effects on problem-solving varied substantially. Among the three studies, only [Samani and Pan \(2021\)](#) found clear interleaving gains on surprise criterial tests made of novel challenging problems. Students improved in the correctness of their solution strategies, and also in long-term memory for relevant formulas. The gain appeared in both halves of the course, and was even larger in the second half, when students had more course material to discriminate against. Interestingly, students rated the interleaved assignments as harder and believed they had learned less, despite consistently performing better. [Gorgievski and DeFranco \(2024\)](#) found no significant difference between blocked and mixed homework on final exam scores.

The reviewed studies point to several implementation factors that appear to be associated with whether interleaving benefits learning: duration and intensity of the intervention, and long-term benefits.

Across these three studies, the most sustained and frequent intervention ([Samani & Pan, 2021](#)) benefited students most clearly. [Samani and Pan's \(2021\)](#) multi-week intervention, with thrice-weekly interleaved assignments, improved performance most. [Gorgievski and DeFranco \(2024\)](#), with a mid-sized sample, ran interleaved homework across a full semester but in a lighter form, mixing problems only within the same chapter rather than across the course. [Gulacar et al. \(2022\)](#), with only three weekend sessions and a small sample, observed a directionally positive but non-significant effect, suggesting that a brief intervention may not be long enough for the two conditions to be detected.

Interleaving benefits appear most clearly on delayed, surprise assessments, and may be masked when students can cram before a test. [Samani and Pan \(2021\)](#) built in a gap of one to three weeks between the last practice and the surprise criterial test. On these delayed surprise tests, the large interleaving benefits emerged. On midterm exams taken a few days later, the same benefits did not appear, because students engaged in extensive cramming between the two events, possibly diminishing the interleaving advantage. [Gorgievski and DeFranco \(2024\)](#) measured performance only on a cumulative semester final exam, with no possibility to control for cramming. [Gulacar et al. \(2022\)](#) used only immediate post-tests and could not measure interleaving's long-term effects.

None of the studies identified student characteristics that reliably moderated the interleaving effect. [Samani and Pan \(2021\)](#) did not formally assess prior knowledge but confirmed comparable GPAs between their counterbalanced sections. [Gorgievski and](#)

DeFranco (2024) used SAT Math scores as a covariate, which correlated with final exam performance and partially controlled for academic background. Gulacar et al. (2022) recorded prior course grades and administered a pretest, then tested whether prior achievement moderated the effect of interleaving. Gulacar et al. (2022) also examined ethnic background as a potential moderator, comparing White, Asian/Pacific, and Latino/Hispanic students. Descriptively, all three groups improved their problem-solving performance with mixed practice, though in different ways. However, subgroup sizes were very small and therefore the effects must be interpreted cautiously.

3.6. Mapping

Mapping is the conversion of linear text into a spatial arrangement of concepts and the links among them, such as a concept map, knowledge map, or graphic organizer (Fiorella & Mayer, 2016).

The evidence for mapping's effectiveness in problem-solving comes from a single, relatively early study, but one that tested performance on novel transfer problems. Okebukola's (1992) research addressed whether concept mapping helps when learning problem-solving. Two groups of biology students, Group A (cooperative) and B (individual), received one week of training in concept mapping, and subsequently created maps during each lesson for six months following a three-step process: listing concepts or main ideas, arranging them hierarchically from abstract to concrete, and adding linking words on connecting lines. Group C (control) received traditional instruction.

Assessment consisted of three novel transfer problems spanning multiple biology domains, requiring students to interrelate concepts rather than recall isolated facts. Students who practiced concept mapping significantly outperformed controls on all problems, with the strongest advantage when problems demanded cross-domain integration. Concept mappers first sketched conceptual relationships before committing to solutions, while control students relied on weak procedures and focused on irrelevant details. This preliminary mapping step reflects the organizing phase of generative learning (Fiorella & Mayer, 2016), thereby developing metacognitive awareness.

Four implementation factors appear particularly relevant. First, the six-month duration allowed sustained practice and integration into learning routines. Second, consistent timing of mapping during each lesson with summary maps every three weeks. Third, problem type moderated effects, with the broadest cross-domain integration showing the largest mapping advantage. Fourth, cooperative versus individual mapping made no significant difference; the benefit resides in the mapping act itself (structuring, hierarchizing, and linking concepts) rather than the social context.

Gender emerged as a significant but inconsistent moderator varying by problem type. Girls outperformed boys on the behaviorally oriented problem, while boys showed stronger performance on the practical laboratory problem. However, girls in both mapping groups consistently outperformed control-group girls across all problems.

3.7. Self-Explanation

Self-explanation requires students to actively explain to themselves the steps or content of instructional material, such as a worked example, thereby constructing meaning that goes beyond what is explicitly stated (Fiorella & Mayer, 2016; Dunlosky et al., 2013).

The five reviewed studies implemented self-explanation in different ways, reflecting how differently self-explanation has been tested across contexts and instructional goals. Four of the five studies used worked examples as the primary instructional material (Badeau et al., 2017; Gjerde et al., 2022; Hausmann & VanLehn, 2010; Miller-Cotto & Auxter, 2021). The one exception, Ionas et al. (2012), embedded self-explanation within direct

problem-solving practice through content-specific guiding questions. Self-explanation was implemented in different forms, such as written free explanations of worked examples (Badeau et al., 2017; Gjerde et al., 2022), verbal self-explanations of video-based worked examples (Hausmann & VanLehn, 2010), answering structured guiding questions before attempting a problem (Ionas et al., 2012), and writing open ‘why’ responses after each problem (Miller-Cotto & Auxter, 2021). Comparison conditions were equally varied, including no training (Badeau et al., 2017, Experiment 1), problem-solving only (Gjerde et al., 2022), paraphrasing (Hausmann & VanLehn, 2010), and business-as-usual problem-solving (Miller-Cotto & Auxter, 2021).

Beyond the form of self-explanation, the studies differed substantially in when and how often students were prompted to self-explain. In most studies, self-explanation was applied immediately and repeatedly during the learning phase: after each segment of the video-based examples (Hausmann & VanLehn, 2010), for each worked example during a single training session (Badeau et al., 2017; Gjerde et al., 2022), or before each problem attempt (Ionas et al., 2012). Miller-Cotto and Auxter (2021) took a more ecological approach, administering self-explanation prompts per problem across four weekly homework assignments over four weeks of instruction.

The studies also diverged in how they measured learning outcomes, capturing how well students transferred and retained what they learned. Badeau et al. (2017) used a novel synthesis problem requiring integration of multiple physics concepts. Gjerde et al. (2022) assessed both problem-solving and conceptual knowledge. Hausmann and VanLehn (2010) measured retention at two points: via an isomorphic homework problem completed shortly after training, and via a far-transfer exam item approximately 29 days later. Ionas et al. (2012) used a single targeted chemistry problem, while Miller-Cotto and Auxter (2021) used the regular classroom exam, making their study the most ecologically valid.

Despite this methodological diversity, these studies point to a consistent pattern: self-explanation indicates an improvement in problem-solving performance, but only under specific conditions. Significant gains were found on complex multi-concept synthesis problems (Badeau et al., 2017), and the quality of students’ explanations was the strongest predictor of performance (Gjerde et al., 2022). Hausmann and VanLehn (2010) found that self-explanation students outperformed students paraphrasing on both training and retention measures, including a far-transfer item 29 days later, supporting the idea that generating content, rather than merely comprehending it, drives learning. By contrast, self-explanation failed to improve and may have harmed performance when prompts were unstructured consisting of a single open ‘why’ question per problem (Miller-Cotto & Auxter, 2021) and was effective only for students above a prior knowledge threshold (Ionas et al., 2012). These boundary conditions suggest five moderating factors: problem complexity (Badeau et al., 2017), self-explanation quality and coverage of principles and conditions (Gjerde et al., 2022), content generation as the active mechanism (Hausmann & VanLehn, 2010), prior knowledge (Ionas et al., 2012), and prompt structure (Miller-Cotto & Auxter, 2021).

The reviewed studies point to several implementation factors that appear to be associated with whether self-explanation benefits learning: the presence of worked examples, the timing of self-explanations, and the prompts’ structure.

How worked examples were structured and how complete they were proved relevant for learning. In four of the five studies, available worked examples served as the material being self-explained (Badeau et al., 2017; Gjerde et al., 2022; Hausmann & VanLehn, 2010; Miller-Cotto & Auxter, 2021), whereas in Ionas et al. (2012) students were asked to self-explain their own reasoning process rather than a pre-solved example. Uniquely among these studies, Hausmann and VanLehn (2010) also manipulated how complete those examples were, comparing self-explanations of fully justified examples against self-explanations

of examples from which justifications had been deliberately omitted. Incomplete examples paired with self-explanation produced the best learning outcomes, outperforming even complete examples paired with self-explanations. This finding supports a generation hypothesis of self-explanation: when students must actively infer missing reasoning steps, the process of generating that content drives deeper learning. In this study, the example of completeness alone, without self-explanation, appeared insufficient.

How often self-explanation was prompted also shaped how effective it was. Most studies embedded self-explanation within a single study session, applying it to each worked example or problem segment in sequence (Badeau et al., 2017; Gjerde et al., 2022; Hausmann & VanLehn, 2010; Ionas et al., 2012). The exception was Miller-Cotto and Auxter (2021), who distributed self-explanations across four weekly homework assignments over four weeks of instruction. Despite the ecological validity of this approach, unstructured self-explanation prompts in that study did not produce benefits over 'business-as-usual' problem-solving, suggesting that self-explanation alone is insufficient without adequate prompt quality.

Among the implementation factors examined here, prompt structure emerges as a particularly influential variable. Self-explanation was delivered verbally (Hausmann & VanLehn, 2010) or in writing (all other studies) and ranged from free explanation to structured prompting. Free written explanations with minimal guidance (Badeau et al., 2017; Gjerde et al., 2022, Study 1) produced mixed but generally positive results, while open 'why' prompts without structure (Miller-Cotto & Auxter, 2021) failed to benefit learners. Gjerde et al. (2022) provided the clearest evidence of this. In Study 1, free written explanations produced variable self-explanation quality. In Study 2, prompts explicitly directed students to articulate principles, conditions of application, and goals; the quality of these explanations was the strongest predictor of performance. Similarly, Ionas et al. (2012) found that structured guiding questions, specifically designed to activate relevant domain knowledge without revealing the answer, were effective above a prior knowledge threshold.

Among student characteristics, prior knowledge may represent an important moderator of self-explanation effectiveness in Ionas et al. (2012). These researchers most directly addressed this variable, demonstrating that the effectiveness of self-explanation prompts on chemistry problem-solving increased with self-reported prior knowledge of the domain. Crucially, below a certain knowledge threshold, self-explanation appears to hinder rather than help performance. Students who lacked the necessary foundation could not construct meaningful self-explanations. Miller-Cotto and Auxter (2021) similarly identified prior knowledge as a moderating variable in developmental algebra, while Gjerde et al. (2022) addressed it implicitly through retrieval practice, on the rationale that self-explanation quality depends on what prior knowledge students can retrieve. Taken together, these findings suggest that self-explanation may amplify what students already know; when that knowledge base is too thin to support meaningful explanation, self-explanation risks becoming counterproductive.

3.8. Testing

Testing, also called retrieval practice, is the use of retrieval as a learning activity, in which learners attempt to produce previously studied material from memory rather than restudying it. Dunlosky et al. (2013) rated practice testing as one of only two learning techniques (alongside distributed practice) at highest utility.

Across nine experiments implementing testing in varied contexts and with different instructional goals, the evidence on problem-solving converges on a clear pattern: structurally embedded testing improves problem-solving, but the effect size depends on what testing is compared with. The most relevant result for novel problem-solving comes from

[Zu et al. \(2019\)](#): retrieval-based practice was marginally better than peer-instructed restudy on delayed and far-transfer tasks. On one side, in a controlled classroom test against passive problem study, retrieval practice produced a very large gain on factual physics knowledge ([Gjerde et al., 2020](#), Phase 1), thus not on problem-solving as such. In intermediate comparisons against traditional assessment schedules, frequent mastery testing with second-chance exams raised final-exam scores and substantially shifted the grade distribution upward ([Morphew et al., 2020](#)), a course-level outcome. On the other side, when retrieval was compared against a similarly active alternative such as peer-instructed restudy ([Zu et al., 2019](#)) or worked examples ([Muzsnay et al., 2024](#)), the testing advantage diminished. Retrieval-based practice was marginally better than peer-instructed restudy and only on delayed and far-transfer tasks ([Zu et al., 2019](#)). Similarly, when researchers compared multiple-attempt practice exams with immediate feedback against a single-attempt format with delayed feedback, exam scores did not change significantly ([M. Zhang et al., 2023](#), Intervention 1). However, three classroom mathematics studies ([May, 2024](#); [Muzsnay et al., 2024](#); [Szabó et al., 2023](#)) reported apparent testing advantages over their business-as-usual comparison conditions.

Implementations followed two main patterns. The most common was to schedule frequent low- or moderate-stakes tests across a full semester and to compare student performance to a prior cohort or a parallel group taught under a traditional assessment schedule ([Bailey et al., 2017](#); [Morphew et al., 2020](#); [Muzsnay et al., 2024](#); [Szabó et al., 2023](#)). By contrast, other studies isolated single-component manipulations: a 70 min retrieval-practice session in introductory mechanics ([Gjerde et al., 2020](#), Phase 1), a single retrieval session following a video lecture ([Zu et al., 2019](#)), and two online practice-exam manipulations within a large physics course, specifically a format change in Intervention 1 and a pre-exam mock in Intervention 2 ([M. Zhang et al., 2023](#)).

The studies also varied in setting, intervention duration, and the delay between testing and the final assessment. Settings included authentic classroom courses ([Bailey et al., 2017](#); [Gjerde et al., 2020](#); [May, 2024](#); [Morphew et al., 2020](#); [Muzsnay et al., 2024](#); [Szabó et al., 2023](#)), a fully online physics course ([M. Zhang et al., 2023](#)), and a research-style classroom session ([Zu et al., 2019](#)). Interventions lasted from a single session ([Gjerde et al., 2020](#), Phase 1; [Zu et al., 2019](#)) to weekly retrieval across the semester ([Bailey et al., 2017](#); [Morphew et al., 2020](#); [Muzsnay et al., 2024](#); [Szabó et al., 2023](#)). Delays varied accordingly, from immediate and 1-week tests ([Zu et al., 2019](#)) to end-of-semester tests ([Bailey et al., 2017](#); [Morphew et al., 2020](#); [Szabó et al., 2023](#)) and a five-month follow-up ([Muzsnay et al., 2024](#)).

The reviewed studies point to several implementation factors that appear to be associated with whether testing benefits learning: structural integration of testing into the course, type and timing of feedback, and spacing and frequency of retrieval opportunities.

Effects tended to be larger when retrieval was a graded part of the course than when it was optional. In [Bailey et al. \(2017\)](#), weekly formative midterms were combined with item-by-item feedback. In [Morphew et al. \(2020\)](#), frequent computer-administered mastery exams with second-chance retakes structurally replaced the two traditional midterms. By contrast, simply offering an optional mock exam, without changing the course's structure, did not change overall student behavior ([M. Zhang et al., 2023](#), Intervention 2). Within the intervention semester, about three-quarters of students opted to complete the mock exam, while the remaining quarter did not. Mock-exam takers were the already-motivated students: they did many practice problems and started practicing earlier than non-takers.

Feedback varied widely and did not follow the rule "more is better". For example, [Morphew et al. \(2020\)](#) combined immediate per-item correctness feedback during the test with a delayed second-chance retake one week later and observed the largest course-level gains in the set. Similarly, [Bailey et al. \(2017\)](#) used posttest in-class feedback from peers and

instructors. However, when M. Zhang et al. (2023, Intervention 1) switched from multiple-attempt practice exams with immediate per-item feedback to single-attempt practice exams with delayed feedback, exam scores did not differ on either hour exam. The new format did not change study behavior, so better feedback alone was not sufficient to improve problem-solving.

Distributing retrieval more frequently was associated with larger and more durable effects. For instance, weekly testing across the semester (Bailey et al., 2017; Morphew et al., 2020) appeared to produce larger and more reliable gains than end-of-class testing (Muzsnay et al., 2024; Szabó et al., 2023), though Muzsnay et al. (2024) reported a long-term retention advantage at five months. Likewise, when retrieval was distributed across many lectures, students who attended more sessions performed better on the end-of-semester exam, especially the weaker students (Gjerde et al., 2020, Phase 2). Single-session retrieval produced an advantage mainly on a delayed assessment one week later (Zu et al., 2019).

Only three studies addressed individual differences directly. First, May (2024) compared students with high and low prior knowledge and reported that low-prior-knowledge students gained more on procedural items while high-prior-knowledge students gained more on conceptual items. However, because sample sizes were small, these effects should be interpreted with caution. Second, Gjerde et al. (2020, Phase 2) used Calculus 1 grade as a proxy for prior ability and observed a significant ability-by-retrieval interaction, with weaker students benefiting more from distributed retrieval and stronger students performing well regardless. Third, Szabó et al. (2023) divided students by initial competence and reported a testing advantage in all three subgroups.

Two studies addressed gender. Szabó et al. (2023) reported no gender effect on the final exam. Likewise, Morphew et al. (2020) found that women, underrepresented minority students, and international students benefited from frequent mastery testing to the same degree as the rest of the sample.

4. Discussion

This systematic literature review examined which desirably difficult and generative learning strategies improve the learning of well-structured problem-solving in STEM higher education, and how this learning is shaped by student characteristics, teacher support, and other implementation factors. Here, we structure the findings around the components of our research question, before reflecting on the alignment between theory and practice.

4.1. Effectiveness of Strategies for Learning Problem-Solving

Our review revealed that desirably difficult and generative strategies improved problem-solving in STEM higher education, but their effects varied in size and robustness depending on the strategy, the subject area, the type of problem, and the design of the intervention.

Across three mathematics studies, distributing practice over weeks produced medium-to-large gains on delayed assessments, and these gains even extended into subsequent courses (Hopkins et al., 2016; Lyle et al., 2020; Y. Zhang et al., 2024). Testing produced benefits across different experiments, with the largest gains in factual physics knowledge (Gjerde et al., 2020) and in courses where retrieval was structurally integrated rather than offered as an option (Bailey et al., 2017; Morphew et al., 2020). For drawing and enacting, the evidence points to an implementation pattern: the benefit lies in active student generation. Students who produced diagrams themselves consistently outperformed students receiving ‘unproductive’ (i.e., instructor-provided) diagrams, and the latter often performed even worse than students who produced no diagram at all (Maries & Singh, 2018a, 2018b, 2023). Similarly, students who produced gestures outperformed students who

observed gestures or received verbal instruction (Stieff et al., 2016a). For self-explanation, benefits were found on complex multi-concept synthesis problems (Badeau et al., 2017) and were strongly correlated with the quality of students' explanations (Gjerde et al., 2022); by contrast, unstructured 'why' questions did not benefit learning and could even harm performance (Miller-Cotto & Auxter, 2021). For interleaving, results were mixed: a sustained frequent intervention produced large gains on surprise tests of novel problems (Samani & Pan, 2021), whereas interventions that were either too brief (three weekend sessions; Gulacar et al., 2022) or too narrow in scope (mix limited to within a single chapter; Gorgievski & DeFranco, 2024) yielded null effects. Mapping evidence rests on a single dated study (Okebukola, 1992) and its results must therefore be treated with caution.

Across these strategies, the most direct evidence for novel or transfer problem gains comes from the few studies in interleaving (Samani & Pan, 2021), concept mapping (Okebukola, 1992), self-explanation (Badeau et al., 2017; Hausmann & VanLehn, 2010), and testing (Zu et al., 2019). Although limited in number, these results provide a strong basis for recommending these strategies when building transferable problem-solving skills.

Beyond the strategy-specific findings, two patterns emerged. The first concerns the timing of assessment: effects on problem-solving increase on delayed assessments and may be invisible on immediate posttests. The second concerns problem complexity: drawing helped most on challenging multi-step problems and was unnecessary for simple ones (Kohl et al., 2007; Maries & Singh, 2018b), and self-explanation benefited synthesis problems more than routine exercises (Badeau et al., 2017). Both patterns are consistent with the principles supporting desirable difficulties (Bjork & Bjork, 2014; Soderstrom & Bjork, 2015): short-term performance and long-term learning are dissociable, and strategies that introduce productive struggle are those whose value only becomes visible over time.

The present review extends the findings of previous literature in several ways. Most outcomes in previous meta-analyses concern retention or comprehension of declarative knowledge, with little evidence on well-structured problem-solving in STEM higher education. For self-explanation, Bisra et al. (2018) report robust benefits across a wide range of subjects and outcomes, with the effect holding across science and mathematics; they locate the mechanism in the extent to which explanation connects new material to what the learner already knows. Our review extends to a problem-solving context and specifies other moderators: prompt structure was decisive, with focused prompts outperforming open ones (Gjerde et al., 2022; Ionas et al., 2012). A minimum threshold of prior knowledge appears necessary for the strategy to be productive (Ionas et al., 2012; Miller-Cotto & Auxter, 2021). For drawing, Leutner and Biele (2025) show, within a generative-learning framework, that drawing benefits learning only when it supports integration of the material. They conclude that it is more effective to support the comprehension process directly than to burden students with additional generative activity that does not serve integration. Our review is consistent with that conclusion and adds specific problem-solving patterns. Ready-made diagrams did not support performance, whereas student-generated diagrams did (Maries & Singh, 2018a, 2018b, 2023). Drawing quality, rather than the mere act of drawing, predicted success (Kohl et al., 2007; Padalkar & Hegarty, 2015). Largest benefits were on difficult problems and when students could compare their own drawing with a correct model (Padalkar & Hegarty, 2015; Stieff et al., 2016b).

4.2. Student Characteristics

Across the studies, prior knowledge was a frequently examined moderator. Several studies suggest that the benefits of generative strategies depend on whether students have enough domain knowledge to generate productive content. Self-explanation was effective only above a prior knowledge threshold and could harm performance below it (Ionas et al.,

2012; Miller-Cotto & Auxter, 2021), and graduate students outperformed undergraduates on diagram-based problem-solving (Maries & Singh, 2023). Testing showed a different pattern: it benefited students across achievement levels, with gains either distributed uniformly (Morphew et al., 2020) or even concentrated among weaker students (Gjerde et al., 2020). Together, these findings suggest that generative strategies such as self-explanation and drawing tend to amplify what students already know and may therefore depend on a minimum threshold of prior knowledge, whereas strategies such as testing and distributed practice appear to benefit students across all levels. This pattern is broadly consistent with the expertise-reversal effect (Kalyuga, 2007) but suggests that not all desirably difficult strategies are equally sensitive to it. Whether these strategies help in practice also depends on whether students adopt them at all: when regulating their own studying, many STEM undergraduates choose low-impact strategies such as re-reading, with strategy use shaped by course design, beliefs, and situation rather than by efficacy (Hora & Oleson, 2017).

Beyond prior knowledge, other student characteristics received much less attention. Gender was examined less systematically, and the effects were small, inconsistent or absent (Hopkins et al., 2016; Morphew et al., 2020; Ping et al., 2022; Szabó et al., 2023). Few studies analyzed ethnicity, motivation, or self-regulation in depth, leaving a substantial gap in our understanding of how student characteristics moderate the effectiveness of these strategies.

4.3. Support and Instructional Design

None of the reviewed studies isolated the teacher's role as an independent variable. No study, for example, contrasted instruction delivered by a teacher with the same instruction delivered without one, nor compared teacher-mediated and computer- or AI-mediated implementations of the same strategy. The broader STEM-education literature indicates that effective implementation depends on instructors' specialized teaching knowledge (Andrews et al., 2022) and on concrete facilitation and explanation strategies that reduce student resistance to the added effort of unfamiliar, active formats (Nguyen et al., 2021). Adoption is, moreover, not only an individual matter: durable use rests on supportive departmental cultures and structures and typically takes years to establish (Lau et al., 2024). The role of the teacher therefore must be inferred indirectly from variations in the instructional design across studies, a limitation we return to later.

Even so, the studies varied considerably in how they scaffolded learning, and four forms of instructional support recurred across these designs. First, scaffolding that involves comparison with a correct model, rather than simply directing students to perform an activity, produced stronger outcomes for drawing and enacting (Padalkar & Hegarty, 2015; Stieff et al., 2016a). When students were simply told to draw, either through detailed or minimal instructions, both approaches produced similar levels of representation use (Kohl et al., 2007). What made the difference was not the strength of the instruction but whether students were given the opportunity to compare their own drawings with a correct model (Stieff et al., 2016b). Second, prompt structure was decisive for self-explanation: prompts directing students to articulate principles or conditions of application supported students to perform better than general open prompts (Gjerde et al., 2022; Jonas et al., 2012). Third, feedback mattered, but more was not always better. Combining item-level correctness feedback during testing with a delayed second-chance retake produced large gains (Morphew et al., 2020). By contrast, changing only the feedback format on practice exams did not alter learning (M. Zhang et al., 2023). Feedback was most effective when students could compare their own output with an external standard (Padalkar & Hegarty, 2015; Stieff et al., 2016a); when feedback only informed without prompting reflection, it added little (M. Zhang et al., 2023). Fourth, worked examples play an important role as a scaffold for self-explanation and as material for students to study. Their effect depends, however,

on how complete they are: Hausmann and VanLehn (2010) showed that worked examples from which justifications had been deliberately omitted produced better learning outcomes than fully elaborated ones, provided that self-explanation was explicitly prompted. When translating these findings to potential teacher support, it appears that setting the conditions under which students engage actively, generate content of sufficient quality, and receive meaningful feedback is central. Future research in this direction is needed.

4.4. Implementation Factors

Beyond strategy choice itself, four implementation factors shaped whether students benefited. First, duration and intensity: longer, more frequent interventions consistently produced more reliable effects than brief ones, particularly for interleaving (Samani & Pan, 2021; Gulacar et al., 2022) and testing (Bailey et al., 2017; Morphew et al., 2020). Second, structural integration: when a strategy was embedded as a graded, mandatory component of the course, it positively influenced all students; when offered optionally, it benefited only those already motivated (M. Zhang et al., 2023). Third, instructional structure: a scaffolded sequence moving from instructor demonstration, through joint performance, to independent student practice characterized all three effective enacting designs (Kadam et al., 2021; Ping et al., 2022; Stieff et al., 2016a). With self-explanation, the most informative worked examples are not necessarily the most complete ones; leaving deliberate gaps for students to fill creates the generative activity that drives learning (Hausmann & VanLehn, 2010). Fourth, setting: laboratory studies tended to produce larger effect sizes on tightly controlled outcomes (e.g., Padalkar & Hegarty, 2015; Y. Zhang et al., 2024), whereas classroom studies produced smaller but more ecologically meaningful effects on real course assessments (e.g., Maries & Singh, 2018b, 2023; Stieff et al., 2016a).

4.5. Theory—Practice Alignment and Ecological Validity

A strength of this body of research is that most studies (25 out of 34) took place in authentic classroom settings, with outcomes measured often on real course exams. This anchors the findings in conditions closer to those in which teachers would apply them. At the same time, laboratory studies often produced larger effect sizes (Padalkar & Hegarty, 2015) than classroom studies on the same strategy (medium-to-large; Maries & Singh, 2018a, 2023; Stieff et al., 2016b), raising the question of how much of the laboratory advantage would survive deployment in a real course. A second consideration is that the assessments used in classroom studies themselves are not always neutral. Several of the largest interleaving (medium-to-large; Samani & Pan, 2021) and distributed-practice (small-to-medium; Y. Zhang et al., 2024) effects emerged only on delayed, surprise assessments and were attenuated or absent on standard midterm exams, where students could compensate through cramming (Samani & Pan, 2021). The choice of assessment can therefore obscure or reveal a strategy's real impact, which complicates the translation of research findings into practice.

A third consideration relates to metacognition. Subjective experiences often diverged from objective gains: students rated interleaved and distributed assignments as harder and believed they had learned less even when they performed better (Samani & Pan, 2021). This metacognitive gap has practical relevance, since it suggests that the perceived difficulty of effective strategies can lead both students and instructors to under-adopt them, in line with the insights from literature (Kirk-Johnson et al., 2019; Baars et al., 2020; Deslauriers et al., 2019; David et al., 2024). Possible interventions to prevent the under-adoption are early instruction on the effort-learning relationship (Deslauriers et al., 2019), or reframing effort as a signal of learning rather than failure (Kirk-Johnson et al., 2019).

The guidance offered above assumes that teachers know these strategies exist and have at least a basic familiarity with the evidence base. In practice, this assumption is often

unwarranted. First, it is uncommon that teachers begin their careers with knowledge of effective learning strategies (Surma et al., 2022). Second, even teachers who can correctly justify the strategies on paper rarely teach them directly in class; classroom observations show indirect modelling vastly exceeding direct instruction (Granström et al., 2023). Third, even when both prerequisites are present, teachers may misjudge the experience of using these strategies; students' struggle and the absence of flow in a lecture can be interpreted as evidence that the strategy is not working. Deslauriers et al. (2019) documented this directly: instructors who tried active strategies often reverted to traditional lectures within a semester, citing student complaints and their own discomfort with the new classroom dynamic. Underlying all three gaps is the fact that literature itself is fragmented across cognitive psychology, discipline-based education research, and instructional design, with terminology that is often inconsistent across fields.

On the assumption that all these conditions can be at least partially met, the present review offers four practical lessons for teachers. First, retrieval and distributed practice should be embedded structurally in courses rather than offered as optional resources. Second, drawing and enacting should be scaffolded through opportunities to compare student work with a correct model. Third, self-explanation prompts should direct attention to underlying principles and conditions of application, rather than asking open 'why' questions. Fourth, worked examples should be designed with deliberate gaps for students to fill. For students, the review highlights one central message: strategies that feel harder and slower in the short term, such as distributed practice, interleaving, retrieval, and generating diagrams or explanations, are precisely those that support long-term problem-solving.

Our introduction framed this review as an attempt to bridge the gap between cognitive science research on effective learning strategies and STEM instructional practice. The present synthesis addresses that gap in three ways: by examining strategy effects on problem-solving rather than on factual retention, by identifying which strategies are most strongly supported as evidence-based practice within STEM higher education and under which conditions, and by making implementation factors explicit so that teachers can adopt strategies in an informed way. Bridging the gap, however, will require complementary implementation research, which we address in the next section.

5. Conclusions, Limitations, and Future Directions

Our introduction identified a persistent gap between cognitive-science research on effective learning strategies and STEM instructional practice, where 'plug-and-chug' routines have remained dominant despite decades of research on other strategies. The present review addresses this gap in three ways. First, it synthesizes evidence specifically for problem-solving outcomes in STEM, most directly for performance on novel problems, rather than for retention of factual material, which is the form in which most prior evidence has been compiled. Second, it identifies which strategies are most defensible as evidence-based practice within STEM higher education, and under which conditions. Third, it makes the implementation factors explicit, so that teachers and curriculum designers can adopt these strategies in informed ways. Closing the gap fully, however, will require complementary implementation research, to which we return below.

There are different limitations in this systematic review. First, our search was limited to three databases, English-language publications, and peer-reviewed journal articles, which excludes potentially relevant work in conference proceedings, dissertations, books, and other languages. Nevertheless, we expect these to be the most relevant publications. Second, we focused exclusively on higher education and on well-structured problems; the findings therefore do not address K-12 contexts, ill-structured problems, or open-ended design tasks that are central in much STEM education. Moreover, most studies enrolled first-year

students, leaving the learning of more advanced students a promising direction for future work. Third, some disciplines (particularly biology and engineering) are represented by only a small number of studies, whereas physics, chemistry, and mathematics dominate the evidence base. This uneven coverage matters because disciplines differ in the kinds of problems they pose, the representations they rely on, and the cognitive demands they place on learners; strategies that work in one context may therefore not transfer straightforwardly to another. Fourth, the risk-of-bias analysis gives some limitations to the conclusions. A few studies have a relatively weak design, mostly small samples without power analyses, and their conclusions were weakened accordingly. This justifies cautious reading. Fifth, the studies varied in how their outcome measures problem-solving: only a few assessed performances on novel problems, while others relied on practiced problems or on knowledge and course-level measures. Finally, several of the originally targeted strategies (elaborative interrogation, summarizing, learning by teaching) yielded no eligible studies, leaving gaps in the synthesis, but also showing potential directions for future research.

These limitations point to directions for future work. First, future research should examine underrepresented disciplines (engineering, biology), underrepresented strategies (mapping, summarizing, teaching), and underexamined moderators (motivation, self-regulation, language background, disciplinary identity). Second, future studies should isolate the teacher's role as an independent variable. At this moment, the contribution of teaching itself remains hidden behind variations in instructional design. Third, researchers should examine individual differences beyond prior knowledge and gender, since motivation and self-regulation remain underexplored, limiting broader generalizability.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/educsci16101733/s1>, Table S1: Details of the experiments on distributed practice included in the systematic review. Table S2: Details of the experiments on drawing included in the systematic review. Table S3: Details of the experiments on enacting included in the systematic review. Table S4: Details of the experiments on interleaving included in the systematic review. Table S5: Details of the experiments on mapping included in the systematic review. Table S6: Details of the experiments on self-explanation included in the systematic review. Table S7: Details of the experiments on testing included in the systematic review.

Author Contributions: Conceptualization, G.T., J.B., F.B., K.B., and A.B.H.d.B.; methodology, G.T., J.B., F.B., K.B., and A.B.H.d.B.; formal analysis, G.T., J.B., and F.B.; investigation, G.T.; resources, G.T.; data curation, G.T., J.B., and F.B.; writing—original draft preparation, G.T.; writing—review and editing, G.T., J.B., F.B., K.B., and A.B.H.d.B.; visualization, G.T.; supervision, K.B., and A.B.H.d.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data and materials used and analyzed for the report are available at various journal websites. We pre-registered our procedure and analysis plan on the Open Science Framework (OSF) website <https://osf.io/hxscd> (pre-registration). Our data set can be found on OSF via <https://osf.io/zh3vr>.

Acknowledgments: We wish to thank Marieke Schor and Rinus Verdonshot, librarians at Maastricht University, for their valuable support during the literature search and screening stages of this review.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Search string used in searching for papers from ERIC database.

Search String Used in Searching for Papers from ERIC Database.	
S1	Topic 1: Problem solving
S2	DE problem solving
S3	TI (problem* N4 solving) OR AB (problem* N4 solving)
	S1 OR S2
S4	Topic 2: STEM
	DE (engineering OR technical education OR Engineering education OR STEM education OR Mathematics education OR Mathematics instruction OR Science education OR Science instruction OR Technology education OR physics OR calculus OR mathematics OR mechanics (physics) OR chemistry OR biology)
S5	TI (stem or science or technology or engineering or mathematics or physics or chemistry or biology) OR AB (stem or science or technology or engineering or mathematics or physics or chemistry or biology)
S6	S4 OR S5
S7	Topic 3: Learning strategies
	DE (teaching methods OR educational strategies OR learning processes OR learning strategies OR instructional design OR mathematics instruction OR science instruction)
S8	TI (*mark* OR "distribut* practice" OR drawing OR elaborate* OR enact* OR "explain* to" OR hierarch* OR highlight* OR imag* OR interleav* OR interrogation OR map OR "mental image" OR "mental repre?entation" OR "mix* practice" OR mnemonic OR organi?er OR quiz* OR recall OR re?reading OR restudy* OR retriev* OR schem* OR "self-expla*" OR spac* OR spreading OR summary* OR teach* OR test* OR underl*) OR AB (*mark* OR "distribut* practice" OR drawing OR elaborate* OR enact* OR "explain* to" OR hierarch* OR highlight* OR imag* OR interleav* OR interrogation OR map OR "mental image" OR "mental repre?entation" OR "mix* practice" OR mnemonic OR organi?er OR quiz* OR recall OR re?reading OR restudy* OR retriev* OR schem* OR "self-expla*" OR spac* OR spreading OR summary* OR teach* OR test* OR underl*)
S9	S7 OR S8
S10	Topic 4: Higher education
	EL ("higher education" or "postsecondary education")
S11	TI ("higher education" or "college*" or "universit*" or "post secondary" or "postsecondary" OR "undergraduate" OR "graduate" OR "postdoctoral" OR "academic" OR "bachelor" OR "polytechnic" OR "masters" OR "post high school" OR "post-secondary" OR "tertiary education" OR "undergraduate") OR AB ("higher education" or "college*" or "universit*" or "post secondary" or "postsecondary" OR "undergraduate" OR "graduate" OR "postdoctoral" OR "academic" OR "bachelor" OR "polytechnic" OR "masters" OR "post high school" OR "post-secondary" OR "tertiary education" OR "undergraduate")
S12	S10 OR S11
S13	S3 AND S6 AND S9 AND S12

Table A2. Search string used in searching for papers from PsycINFO database.

Search String Used in Searching for Papers from PsycINFO Database.	
S1	Topic 1: Problem solving
S2	DE problem solving
S3	TI (problem* N4 solving) OR AB (problem* N4 solving)
	S1 OR S2
S4	Topic 2: STEM
	DE (engineering OR STEM OR Sciences OR Mathematics education OR Science education OR physics OR calculus OR mathematics OR chemistry OR biology)
S5	TI (stem or science or technology or engineering or mathematic* or physics or chemistry or biology) OR AB (stem or science or technology or engineering or mathematic* or physics or chemistry or biology)
S6	S4 OR S5

Table A2. Cont.

Search String Used in Searching for Papers from PsycINFO Database.	
S7	Topic 3: Learning strategies DE ("Teaching methods" OR "Heuristics" OR "Learning" OR "Learning strategies" OR Teaching)
S8	TI (*mark* OR "distribut* practice" OR drawing OR elaborate* OR enact* OR "explain* to" OR hierarch* OR highlight* OR imag* OR interleav* OR interrogation OR map OR "mental image" OR "mental repre?entation" OR "mix* practice" OR mnemonic OR organi?er OR quiz* OR recall OR re*reading OR restudy* OR retriev* OR schem* OR "self-expla*" OR spac* OR spreading OR summary* OR teach* OR test* OR underl*) OR AB (*mark* OR "distribut* practice" OR drawing OR elaborate* OR enact* OR "explain* to" OR hierarch* OR highlight* OR imag* OR interleav* OR interrogation OR map OR "mental image" OR "mental repre?entation" OR "mix* practice" OR mnemonic OR organi?er OR quiz* OR recall OR re*reading OR restudy* OR retriev* OR schem* OR "self-expla*" OR spac* OR spreading OR summary* OR teach* OR test* OR underl*)
S9	S7 OR S8
S10	Topic 4: Higher education DE ("College Students" OR "Higher Education" OR "Graduate Education" OR "Graduate Students" OR "Postgraduate Training" OR "Undergraduate Education" OR "Colleges")
S11	TI ("higher education" or "college*" or "universit*" or "post secondary" or "postsecondary" OR "undergraduate" OR "graduate" OR "postdoctoral" OR "academic" OR "bachelor" OR "polytechnic" OR "masters" OR "post high school" OR "post-secondary" OR "tertiary education" OR "undergraduate") OR AB ("higher education" or "college*" or "universit*" or "post secondary" or "postsecondary" OR "undergraduate" OR "graduate" OR "postdoctoral" OR "academic" OR "bachelor" OR "polytechnic" OR "masters" OR "post high school" OR "post-secondary" OR "tertiary education" OR "undergraduate")
S12	S10 OR S11
S13	S3 AND s6 AND S9 AND S12

Table A3. Search string used in searching for papers from Web of Science database.

Search String Used in Searching for Papers from Web of Science Database.	
#1	Topic 1: Problem solving TS = (problem* NEAR/4 solv*)
#2	Topic 2: STEM TS = (stem or science or technology or engineering or mathematics or physics or chemistry or biology)
#3	Topic 3: Learning strategies TS = (mark* OR unmark* OR "distribut* practice" OR drawing OR elaborate* OR enact* OR "explain* to" OR hierarch* OR highlight* OR imag* OR interleav* OR interrogation OR map OR "mental image" OR "mental repre\$entation" OR "mix* practice" OR mnemonic OR organi\$er OR quiz* OR recall OR re*reading OR restudy* OR retriev* OR schem* OR "self-expla*" OR spac* OR spreading OR summary* OR teach* OR test* OR underl*)
#4	Topic 4: Higher education TS = ("higher education" or "college*" or "universit*" or "post secondary" or "postsecondary" OR "undergraduate" OR "graduate" OR "postdoctoral" OR "academic" OR "bachelor" OR "polytechnic" OR "masters" OR "post high school" OR "post-secondary" OR "tertiary education" OR "undergraduate")
#5	#1 AND #2 AND #3 AND #4

Appendix B

Table A4. Risk-of-bias analysis for the studies included in the systematic review.

Study	Research Question or Aim of Study	Theoretical or Conceptual Underpinning	Participants	Power	Data Collection Method	Analysis of the Results	Reporting Results	Reproducibility	Strengths and Limitations	Conclusions	Overall Result
	Is the research question(s) or aim of the study clearly stated?	Are specific theories or concepts that frame the study identified?	Is the subject group appropriate for the study being carried out (number, characteristics, selection, and homogeneity)?	Did the study have sufficient power to detect an important effect?	Are the methods used (qualitative or quantitative) reliable and valid for the research question and context?	Are the statistical or other methods of results analysis used appropriate?	Are the main findings of the study clearly described?	Could the study be repeated by other researchers?	Are the key strengths and weaknesses of the study critically discussed?	Is it clear that the data justify the conclusions drawn?	
Badeau et al. (2017)	1	1	1	1	1	1	1	1	1	1	10
Bahtaji (2020)	1	1	1	0	0	1	1	0	0	0	5
Bailey et al. (2017)	1	1	0	0	1	1	1	1	0	0	6
Gjerde et al. (2022)	1	1	1	0	1	1	1	1	1	1	9
Gjerde et al. (2020)	1	1	1	1	1	1	1	1	1	1	10
Gorgievski and DeFranco (2024)	1	1	1	0	1	1	1	1	0	1	8
Gulacar et al. (2022)	1	1	0	0	1	1	1	0	1	0	6
Hausmann and VanLehn (2010)	1	1	1	1	1	1	1	1	1	1	10
Heckler (2010)	1	1	1	1	1	1	1	1	0	1	9
Heideman et al. (2017)	1	1	0	0	1	1	1	1	1	1	8
Hopkins et al. (2016)	1	1	1	1	1	1	1	1	1	1	10
Jonas et al. (2012)	1	1	1	0	0	1	1	0	1	1	7
Kadam et al. (2021)	1	1	1	0	1	1	1	1	0	0	7
Kohl et al. (2007)	1	1	1	1	1	1	1	1	1	1	10
Lyle et al. (2020)	1	1	1	1	1	1	1	1	1	1	10
Lyle et al. (2022)	1	1	1	1	1	1	1	1	1	1	10
Maries and Singh (2018a)	1	1	1	0	1	1	1	1	0	1	8
Maries and Singh (2018b)	1	1	1	0	1	1	1	1	0	1	8
Maries and Singh (2023)	1	1	1	0	1	1	1	1	0	1	8
May (2024)	1	1	0	0	0	1	1	1	1	0	6
Miller-Cotto and Auxter (2021)	1	1	1	0	1	1	1	1	1	1	9
Morphew et al. (2020)	1	1	1	1	1	1	1	1	1	1	10
Muzsnay et al. (2024)	1	1	1	0	1	1	1	1	1	1	9
Okebukola (1992)	1	1	1	0	1	1	1	0	0	0	6
Padalkar and Hegarty (2015)	1	1	1	0	1	1	1	1	1	1	9
Ping et al. (2022)	1	1	1	1	1	1	1	1	1	1	10
Samani and Pan (2021)	1	1	1	1	1	1	1	1	1	1	10

Table A4. Cont.

Study	Research Question or Aim of Study	Theoretical or Conceptual Underpinning	Participants	Power	Data Collection Method	Analysis of the Results	Reporting Results	Reproducibility	Strengths and Limitations	Conclusions	Overall Result
Stieff et al. (2016a)	1	1	1	0	1	1	1	1	1	1	9
Stieff et al. (2016b)	1	1	1	1	1	1	1	1	1	1	10
Szabó et al. (2023)	1	1	1	0	1	1	1	1	1	1	9
Wu et al. (2020)	1	1	1	1	1	1	1	1	1	1	10
M. Zhang et al. (2023)	1	1	1	1	1	1	1	1	1	1	10
Y. Zhang et al. (2024)	1	1	1	0	1	1	1	1	1	1	9
Zu et al. (2019)	1	1	1	0	1	1	1	1	1	1	9

Note. The initial appraisal was conducted by the first author. The studies with an uncertain appraisal were independently evaluated by the second author against the source articles. Disagreements were resolved by discussion.

References

- Adams, W. K., & Wieman, C. E. (2015). Analyzing the many skills involved in solving complex physics problems. *American Journal of Physics*, 83(5), 459–467. [\[CrossRef\]](#)
- Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, 87(3), 659–701. [\[CrossRef\]](#)
- Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*, 33(4), 1409–1453. [\[CrossRef\]](#)
- Anderson, J. R. (1980). *Cognitive psychology and its implications*. Freeman.
- Andrews, T. C., Speer, N. M., & Shultz, G. V. (2022). Building bridges: A review and synthesis of research on teaching knowledge for undergraduate instruction in science, engineering, and mathematics. *International Journal of STEM Education*, 9(1), 66. [\[CrossRef\]](#)
- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence, & J. T. Spence (Eds.), *Psychology of learning and motivation* (Vol. 2, pp. 89–195). Academic Press. [\[CrossRef\]](#)
- Baars, M., Wijnia, L., de Bruin, A., & Paas, F. (2020). The relation between students' effort and monitoring judgments during learning: A meta-analysis. *Educational Psychology Review*, 32(4), 979–1002. [\[CrossRef\]](#)
- Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 8, pp. 47–89). Academic Press. [\[CrossRef\]](#)
- Badeau, R., White, D. R., Ibrahim, B., Ding, L., & Heckler, A. F. (2017). What works with worked examples: Extending self-explanation and analogical comparison to synthesis problems. *Physical Review Physics Education Research*, 13(2), 020112. [\[CrossRef\]](#)
- Bahtaji, M. A. A. (2020). Improving students graphing skills and conceptual understanding using explicit graphical physics instructions. *Cypriot Journal of Educational Sciences*, 15(4), 843–853. [\[CrossRef\]](#)
- Bailey, E. G., Jensen, J., Nelson, J., Wiberg, H. K., Bell, J. D., & Marcey, D. (2017). Weekly formative exams and creative grading enhance student learning in an introductory biology course. *CBE Life Sciences Education*, 16(1), ar2. [\[CrossRef\]](#) [\[PubMed\]](#)
- Bergqvist, E. (2007). Types of reasoning required in university exams in mathematics. *The Journal of Mathematical Behavior*, 26(4), 348–370. [\[CrossRef\]](#)
- Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, 30(3), 703–725. [\[CrossRef\]](#)
- Bjork, E. L., & Bjork, R. A. (2014). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, & J. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (2nd ed., pp. 59–68). Worth Publishers.
- Boetje, J., & van de Schoot, R. (2024). The SAFE procedure: A practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. *Systematic Reviews*, 13(1), 81. [\[CrossRef\]](#) [\[PubMed\]](#)
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. [\[CrossRef\]](#)
- Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, 145(11), 1029–1052. [\[CrossRef\]](#) [\[PubMed\]](#)
- Buckley, S., Coleman, J., Davison, I., Khan, K. S., Zamora, J., Malick, S., Morley, D., Pollard, D., Ashcroft, T., Popovic, C., & Sayers, J. (2009). The educational effects of portfolios on undergraduate student learning: A best evidence medical education (BEME) systematic review. BEME guide no. 11. *Medical Teacher*, 31(4), 282–298. [\[CrossRef\]](#) [\[PubMed\]](#)
- Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, 1(9), 496–511. [\[CrossRef\]](#)
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2), 121–152. [\[CrossRef\]](#) [\[PubMed\]](#)
- David, L., Vassena, E., & Bijleveld, E. (2024). The unpleasantness of thinking: A meta-analytic review of the association between mental effort and negative affect. *Psychological Bulletin*, 150(9), 1070–1093. [\[CrossRef\]](#) [\[PubMed\]](#)
- Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39), 19251–19257. [\[CrossRef\]](#) [\[PubMed\]](#)
- Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health*, 52(6), 377. [\[CrossRef\]](#) [\[PubMed\]](#)
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, 14(1), 4–58. [\[CrossRef\]](#) [\[PubMed\]](#)
- Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, 28(4), 717–741. [\[CrossRef\]](#)

- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415. [\[CrossRef\]](#) [\[PubMed\]](#)
- Gick, M. L. (1986). Problem-solving strategies. *Educational Psychologist*, 21(1–2), 99–120. [\[CrossRef\]](#)
- Gjerde, V., Havre Paulsen, V., Holst, B., & Kolstø, S. D. (2022). Problem solving in basic physics: Effective self-explanations based on four elements with support from retrieval practice. *Physical Review Physics Education Research*, 18(1), 010136. [\[CrossRef\]](#)
- Gjerde, V., Holst, B., & Kolstø, S. D. (2020). Retrieval practice of a hierarchical principle structure in university introductory physics: Making stronger students. *Physical Review Physics Education Research*, 16(1), 013103. [\[CrossRef\]](#)
- Gomes, S. R., Saroar, S. G., Mosfaiul, M., Telot, A., Khan, B. N., Chakrabarty, A., & Mostakim, M. (2017). A comparative approach to email classification using Naive Bayes classifier and hidden Markov model. In *4th international conference on advances in electrical engineering (ICAEE)*. IEEE.
- Gorgievski, N., & DeFranco, T. (2024). The impact of blocked practice versus mixed practice and the strategy of overlearning on student performance in calculus. *Investigations in Mathematics Learning*, 16(2), 137–151. [\[CrossRef\]](#)
- Granström, M., Kikas, E., & Eisenschmidt, E. (2023). Classroom observations: How do teachers teach learning strategies? *Frontiers in Education*, 8, 1119519. [\[CrossRef\]](#)
- Gulacar, O., Wu, A., Prathikanti, V., Vernoy, B., Kim, H., Bacha, T., Oentoro, T., Navarrete-Pleitez, M., & Reedy, K. (2022). Benefits of desirable difficulties: Comparing the influence of mixed practice to that of categorized sets of questions on students' problem-solving performance in chemistry. *Chemistry Education Research and Practice*, 23(2), 422–435. [\[CrossRef\]](#)
- Haddaway, N. A.-O., Grainger, M. A.-O., & Gray, C. T. (2022). Citationchaser: A tool for transparent and efficient forward and backward citation chasing in systematic searching. *Research Synthesis Methods*, 13(4), 533–545. [\[CrossRef\]](#) [\[PubMed\]](#)
- Harrison, R., Jones, B., Gardner, P., & Lawton, R. (2021). Quality assessment with diverse studies (QuADS): An appraisal tool for methodological and reporting quality in systematic reviews of mixed- or multi-method studies. *BMC Health Services Research*, 21(1), 144. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hausmann, R. G. M., & VanLehn, K. (2010). The effect of self-explaining on robust learning. *International Journal of Artificial Intelligence in Education*, 20(4), 303–332. [\[CrossRef\]](#)
- Heckler, A. F. (2010). Some consequences of prompting novice physics students to construct force diagrams. *International Journal of Science Education*, 32(14), 1829–1851. [\[CrossRef\]](#)
- Heideman, P. D., Flores, K. A., Sevier, L. M., & Trouton, K. E. (2017). Effectiveness and adoption of a drawing-to-learn study tool for recall and problem solving: Minute sketches with folded lists. *CBE—Life Sciences Education*, 16(2), 28. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hopkins, R. F., Lyle, K. B., Hieb, J. L., & Ralston, P. A. (2016). Spaced retrieval practice increases college students' short- and long-term retention of mathematics knowledge. *Educational Psychology Review*, 28(4), 853–873. [\[CrossRef\]](#)
- Hora, M. T., & Oleson, A. K. (2017). Examining study habits in undergraduate STEM courses from a situative perspective. *International Journal of STEM Education*, 4(1), 1. [\[CrossRef\]](#)
- Ibrahim, B., Ding, L., Heckler, A. F., White, D. R., & Badeau, R. (2017). Students' conceptual performance on synthesis physics problems with varying mathematical complexity. *Physical Review Physics Education Research*, 13(1), 010133. [\[CrossRef\]](#)
- Ionas, I. G., Cernusca, D., & Collier, H. L. (2012). Prior knowledge influence on self-explanation effectiveness when solving problems: An exploratory study in science learning. *International Journal of Teaching and Learning in Higher Education*, 24(3), 349–358.
- Jackson, P., & Moulinier, I. (2007). *Natural language processing for online applications*. John Benjamins. [\[CrossRef\]](#)
- Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, 48(4), 63–85. [\[CrossRef\]](#)
- Kadam, K., Mishra, S., Deep, A., & Iyer, S. (2021). Enhancing engineering drawing skills via fostering mental rotation processes. *European Journal of Engineering Education*, 46(5), 796–812. [\[CrossRef\]](#)
- Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, 19(4), 509–539. [\[CrossRef\]](#)
- Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, 115, 101237. [\[CrossRef\]](#) [\[PubMed\]](#)
- Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19(3), 239–264. [\[CrossRef\]](#)
- Kohl, P. B., Rosengrant, D., & Finkelstein, N. D. (2007). Strongly and weakly directed approaches to teaching multiple representation use in physics. *Physical Review Special Topics—Physics Education Research*, 3(1), 010108. [\[CrossRef\]](#)
- Lau, A. C., Henderson, C., Stains, M., Dancy, M., Merino, C., Apkarian, N., Raker, J. R., & Johnson, E. (2024). Characteristics of departments with high-use of active learning in introductory STEM courses: Implications for departmental transformation. *International Journal of STEM Education*, 11(1), 10. [\[CrossRef\]](#)

- Leutner, D., & Biele, J. (2025). Without integration, everything is nothing: A meta-analysis of the effectiveness of instructional support for drawing-to-learn. *Educational Psychology Review*, 37(4), 93. [\[CrossRef\]](#)
- Lima, R. M., Andersson, P. H., & Saalman, E. (2017). Active learning in engineering education: A (re)introduction. *European Journal of Engineering Education*, 42(1), 1–4. [\[CrossRef\]](#)
- Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. S. (2020). How the amount and spacing of retrieval practice affect the short- and long-term retention of mathematics knowledge. *Educational Psychology Review*, 32(1), 277–295. [\[CrossRef\]](#)
- Lyle, K. B., Bego, C. R., Ralston, P. A. S., & Immekus, J. C. (2022). Spaced retrieval practice imposes desirable difficulty in calculus learning. *Educational Psychology Review*, 34(3), 1799–1812. [\[CrossRef\]](#)
- Maries, A., & Singh, C. (2018a). Case of two electrostatics problems: Can providing a diagram adversely impact introductory physics students' problem solving performance? *Physical Review Physics Education Research*, 14(1), 010114. [\[CrossRef\]](#)
- Maries, A., & Singh, C. (2018b). Do students benefit from drawing productive diagrams themselves while solving introductory physics problems? The case of two electrostatics problems. *European Journal of Physics*, 39(1), 015703. [\[CrossRef\]](#)
- Maries, A., & Singh, C. (2023). Helping students become proficient problem solvers part II: An example from waves. *Education Sciences*, 13(2), 138. [\[CrossRef\]](#)
- May, B. M. (2024). Retrieval practice and test-potentiated learning: A comparison of high and low prior topic knowledge students in terms of procedural fluency and conceptual understanding. *Pythagoras*, 45(1), 14. [\[CrossRef\]](#)
- Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). Cambridge University Press. [\[CrossRef\]](#)
- Miller-Cotto, D., & Auxter, A. E. (2021). Testing the ecological validity of faded worked examples in algebra. *Educational Psychology*, 41(2), 191–205. [\[CrossRef\]](#)
- Morphew, J. W., Silva, M., Herman, G., & West, M. (2020). Frequent mastery testing with second-chance exams leads to enhanced student learning in undergraduate engineering. *Applied Cognitive Psychology*, 34(1), 168–181. [\[CrossRef\]](#)
- Muzsnay, A., Zámbo, C., Szeibert, J., Bernáth, L., Szilágyi, B., & Szabó, C. (2024). How do testing and test-potentiated learning versus worked example method affect medium- and long-term knowledge in abstract algebra for pre-service mathematics teachers? *European Journal of Psychology of Education*, 39(4), 4395–4415. [\[CrossRef\]](#)
- Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, 76(3), 413–448. [\[CrossRef\]](#)
- Newell, H. A., & Simon, A. (1972). *Human problem solving*. Prentice Hall.
- Nguyen, K. A., Borrego, M., Finelli, C. J., DeMonbrun, M., Crockett, C., Tharayil, S., Shekhar, P., Waters, C., & Rosenberg, R. (2021). Instructor strategies to aid implementation of active learning: A systematic literature review. *International Journal of STEM Education*, 8(1), 9. [\[CrossRef\]](#)
- Okebukola, P. A. (1992). Can good concept mappers be good problem solvers in science? *Educational Psychology*, 12(2), 113–129. [\[CrossRef\]](#)
- Padalkar, S., & Hegarty, M. (2015). Models as feedback: Developing representational competence in chemistry. *Journal of Educational Psychology*, 107(2), 451–467. [\[CrossRef\]](#)
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. [\[CrossRef\]](#) [\[PubMed\]](#)
- Ping, R., Parrill, F., Church, R. B., & Goldin-Meadow, S. (2022). Teaching stereoisomers through gesture, action, and mental imagery. *Chemistry Education Research and Practice*, 23(3), 698–713. [\[CrossRef\]](#)
- Polya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press. [\[CrossRef\]](#)
- Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using siamese BERT-networks*. Association for Computational Linguistics. [\[CrossRef\]](#)
- Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology*, 91(1), 175–189. [\[CrossRef\]](#)
- Roediger, H. L., III, & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition*, 1(4), 242–248. [\[CrossRef\]](#)
- Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology*, 20(9), 1209–1224. [\[CrossRef\]](#)
- Ruitenburt, S. K., Ackermans, K., Kirschner, P. A., Jarodzka, H., & Camp, G. (2025). After initial acquisition, problem-solving leads to better long-term performance than example study, even for complex tasks. *Learning and Instruction*, 95, 102027. [\[CrossRef\]](#)
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. [\[CrossRef\]](#)

- Samani, J., & Pan, S. C. (2021). Interleaved practice enhances memory and problem-solving ability in undergraduate physics. *npj Science of Learning*, 6(1), 32. [\[CrossRef\]](#) [\[PubMed\]](#)
- Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2), 176–199. [\[CrossRef\]](#) [\[PubMed\]](#)
- Stieff, M., Lira, M. E., & Scopelitis, S. A. (2016a). Gesture supports spatial thinking in STEM. *Cognition and Instruction*, 34(2), 80–99. [\[CrossRef\]](#)
- Stieff, M., Scopelitis, S., Lira, M. E., & DeSutter, D. (2016b). Improving representational competence with concrete models. *Science Education*, 100(2), 344–363. [\[CrossRef\]](#)
- Surma, T., Camp, G., de Groot, R., & Kirschner, P. A. (2022). Novice teachers' knowledge of effective study strategies. *Frontiers in Education*, 7, 996039. [\[CrossRef\]](#)
- Szabó, C., Bereczky-Zámbó, C., Muzsnay, A., Szeibert, J., & Bernáth, L. (2023). Investigating the efficacy of retrieval practice in university mathematics. *Revista de Educación*, 401, 79–96. [\[CrossRef\]](#)
- Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintrón, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones, L., Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., . . . Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences*, 117(12), 6476–6483. [\[CrossRef\]](#) [\[PubMed\]](#)
- Van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdemá, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2), 125–133. [\[CrossRef\]](#)
- Walsh, L. N., Howard, R. G., & Bowe, B. (2007). Phenomenographic study of students' problem solving approaches in physics. *Physical Review Special Topics—Physics Education Research*, 3(2), 020108. [\[CrossRef\]](#)
- Wittrock, M. C. (1974). Learning as a generative process 1. *Educational Psychologist*, 11(2), 87–95. [\[CrossRef\]](#)
- Wu, S. P. W., Van Veen, B., & Rau, M. A. (2020). How drawing prompts can increase cognitive engagement in an active learning engineering course. *Journal of Engineering Education*, 109(4), 723–742. [\[CrossRef\]](#)
- Yu, Z., Kraft, N. A., & Menzies, T. (2018). Finding better active learners for faster literature reviews. *Empirical Software Engineering*, 23(6), 3161–3186. [\[CrossRef\]](#)
- Zhang, M., Morphey, J., & Stelzer, T. (2023). Impact of more realistic and earlier practice exams on student metacognition, study behaviors, and exam performance. *Physical Review Physics Education Research*, 19(1), 010130. [\[CrossRef\]](#)
- Zhang, Q., & Fiorella, L. (2021). Learning by drawing: When is it worth the time and effort? *Contemporary Educational Psychology*, 66, 101990. [\[CrossRef\]](#)
- Zhang, Y., Wang, Z., Chen, Z., Yang, Q., Dong, C., & Qu, K. (2024). Spaced example-based learning facilitates advanced mathematics learning in college students. *Journal of Cognitive Psychology*, 36(2), 205–219. [\[CrossRef\]](#)
- Zu, T., Munsell, J., & Rebello, N. S. (2019). Comparing retrieval-based practice and peer instruction in physics learning. *Physical Review Physics Education Research*, 15(1), 010105. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.