A systematic review and meta-analysis of interventions based on metacognition and self-regulation in school-aged mathematics

HITT, LORAINE,ELISABETH

Author: Loraine Elisabeth Hitt

Title: A systematic review and meta-analysis of interventions based on metacognition and self-regulation in school-aged mathematics

Abstract: Mathematics is an important gatekeeper for educational and professional opportunities and a useful tool for discovery and expression. Given previous research and theory demonstrating potential for metacognitive and self-regulated learning (MC/SRL) interventions to support mathematics achievement with diverse learners, a systematic review was conducted to evaluate their effectiveness within the years of general education, with pupils of ages three to 18. Appropriately-designed studies that were reported in English between 2005 and 2019 were included. Following a systematic search, with double-reviewing and expert consultation for consistency, 1,761 bibliographic items were screened, resulting in 60 included studies. Qualitative aspects of the designs, contexts, participants, and intervention activities were synthesised narratively. Posttest-only and adjusted, random effects meta-analyses were performed using a single mathematics achievement measure from each study. The results indicate a generally positive effect from the included interventions (combined Cohen's *d*=0.46, SE=0.08, 95% CI=0.30 to 0.60). This represents a somewhat more modest effect compared with previous reviews in this area, possibly due to a greater range of included reports. No risk of publication bias was identified, reflecting the breadth and diversity of included studies, but efforts to mitigate heterogeneity were only partially successful. Interventions using structured problem-solving with metacognitive prompts were more effective than those not using it, while dissertations reported lower effects than journal articles. No differences were found based on participant age or intervention dose. Primary studies used a variety of assessments and differed on reporting of interventions and quality-related factors, and there remained substantial heterogeneity in the meta-analysis. Implications of this review for educational theory, research, and practice are discussed, with emphasis on reporting studies fully, using broad-scope, comparable assessments, and investing in comprehensive metacognitive and self-regulated learning interventions that can support lasting change in teaching and learning.

A systematic review and meta-analysis of interventions based on metacognition and self-regulation in school-aged mathematics

by

Loraine Elisabeth Hitt

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Education

Durham University

School of Education

2023

Table of Contents

List of Tables

List of Figures

Statement of Copyright

Acknowledgements

Throughout my doctoral study, I have been blessed with the constant support and encouragement of my family, especially Samuel, Caleb, and my husband, Wes. I also owe a deep debt of gratitude to my instructors, supervisors, and colleagues in the School of Education and across Durham University. Your expertise, hard work, graciousness, and commitment to improving education are inspiring, and it has been a privilege to work alongside you these past years. I am so grateful for the connections we have made and for the chance to demonstrate returns on the investments you have made in me. A special acknowledgement is due for the invaluable guidance and support of my first supervisor, Professor Steve Higgins, who embodies all the qualities I hope to have someday as an education researcher and mentor. Professor Julie Rattray read and gave encouraging feedback on the thesis, and my mother, Sherry Kuyt, provided invaluable copyediting support. I would also like to acknowledge and thank the support staff at Durham and at EPPI-Reviewer, who have been unfailingly responsive and helpful in times of need. Finally, I would like to acknowledge the contributions of the students, teachers, and researchers represented by the hundreds of education studies I encountered in this review. Thank you for your contributions to improving teaching and learning outcomes for future students around the world.

Dedication

This thesis is lovingly dedicated to my family, and to the glory of God.

> For God, who commanded the light to shine out of darkness, hath shone in our hearts, to give the light of the knowledge of the glory of God in the face of Jesus Christ. But we have this treasure in earthen vessels, that the excellency of the power may be of God, and not of us. (II Corinthians 4:6-7)

Chapter 1: Introduction

1.1 The research context and rationale

Modern societies value measurable results, and this is especially true in the context of learning in school. The project of universal education has been largely triumphant in raising literacy (Figure 1) and numeracy rates (Figure 2) and economic opportunities for students around the world to unprecedented levels, but not all pupils have benefited equally. In fact, the most privileged students have consistently demonstrated the highest school achievement levels, and it is not clear that disparities in outcomes are improving (Schleicher, 2019, p. 5). During the recent pandemic, school closures put pressure on already over-taxed educational systems, with many localities attempting and failing to provide high-quality distance-learning opportunities. We are still coming to grips with the academic impacts, with especially large gaps in mathematics and for poorer and minority students (Kuhfeld & Lewis, 2022, pp. 5-7). Even prior to COVID-19, there was a sense that students needed to be prepared for an increasingly complex working world, with more frequent changes in work environments and required skills, especially in light of new technologies. Despite geopolitical tensions, global mobility has never been higher, and the need to compete for opportunities is frequently expressed. TIMMS and PISA rankings have been cited to demonstrate the ascendancy or decline of nations relative to each other (e.g., Wright, 2022), as much as to show the improvement or maintenance of educational quality within a single nation. The perception is that the stakes for education grow ever higher, while achievement itself (and possibly the standards by which it is judged) stagnate.

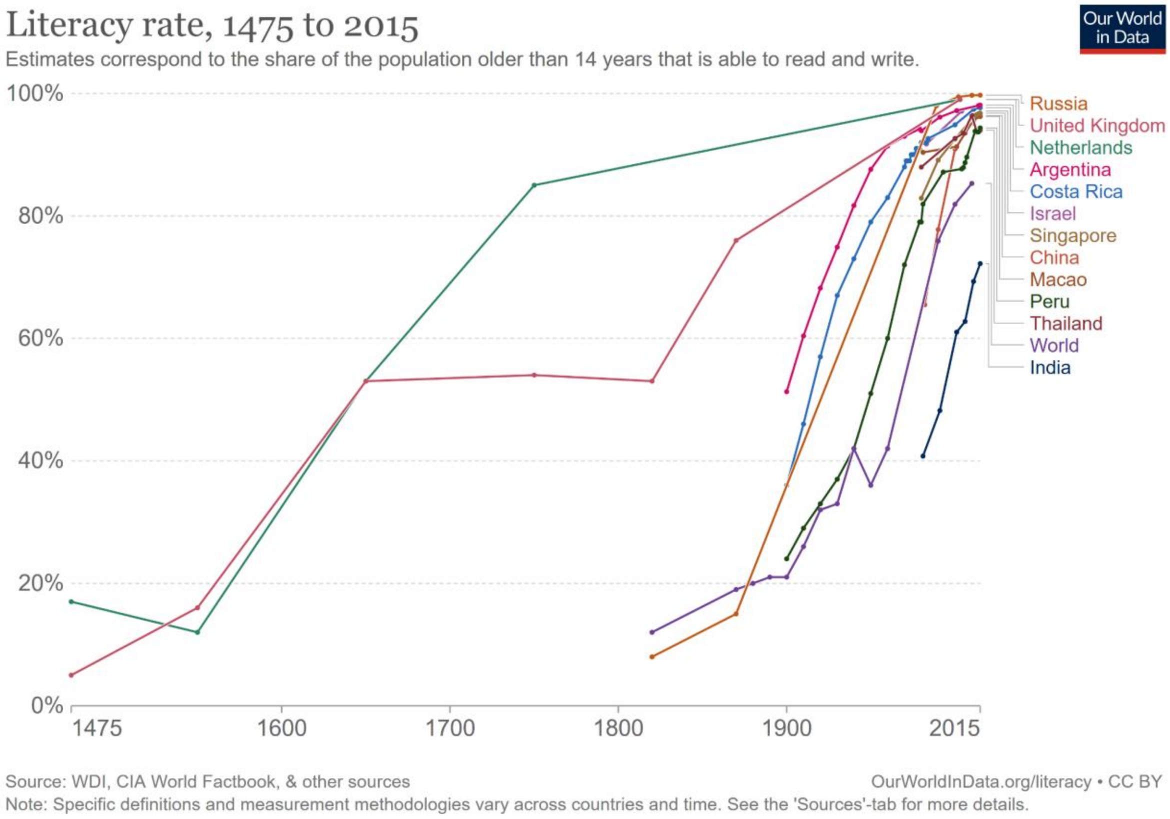
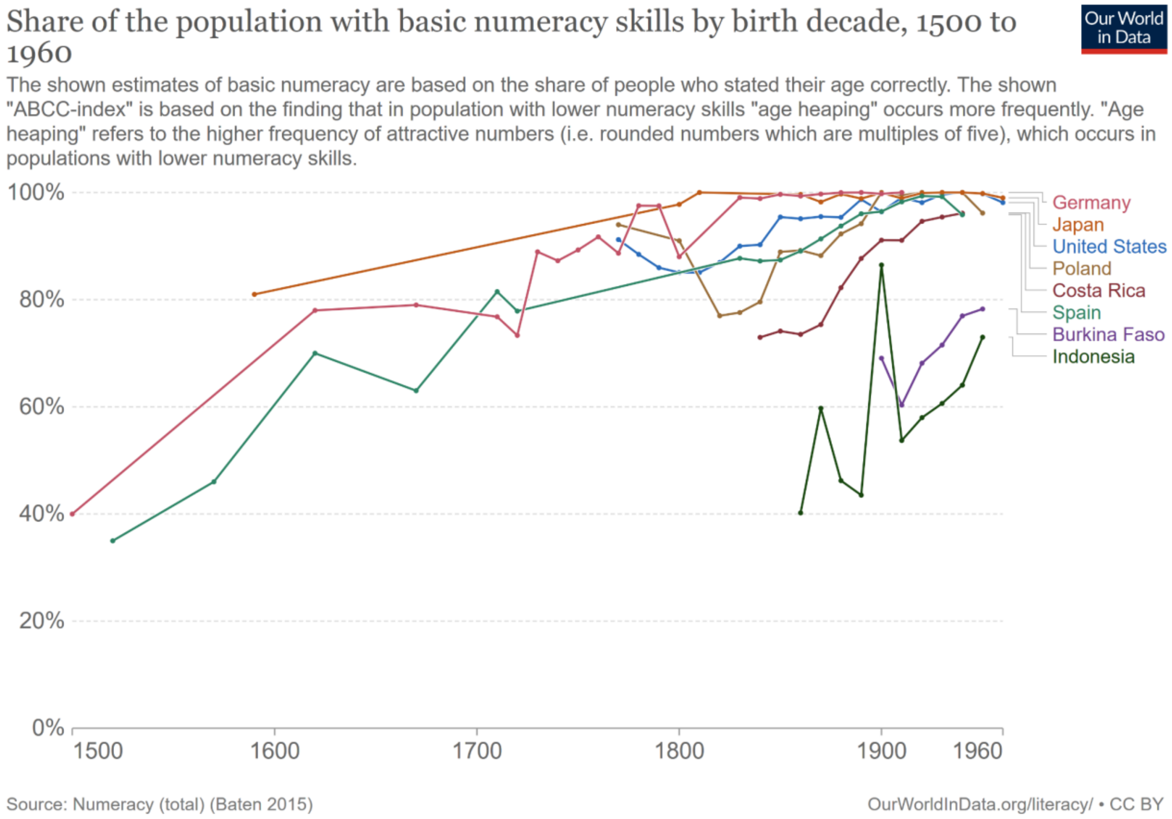


Figure 1. Literacy rates by country and year.
Graph reused from Roser & Ortiz-Ospina (2016), under "creative commons."



Share of the population with basic numeracy skills by birth decade, 1500 to 1960

The shown estimates of basic numeracy are based on the share of people who stated their age correctly. The shown "ABCC-index" is based on the finding that in population with lower numeracy skills "age heaping" occurs more frequently. "Age heaping" refers to the higher frequency of attractive numbers (i.e. rounded numbers which are multiples of five), which occurs in populations with lower numeracy skills.

Source: Numeracy (total) (Baten 2015)

OurWorldInData.org/literacy/ • CC BY

Figure 2. Basic numeracy rates by country and year.
Graph reused from Roser & Ortiz-Ospina (2016), under "creative commons."

Within this context, an explicit focus on soft skills, such as metacognition and self-regulated learning, could seem like a distraction. Every minute students spend writing journals, discussing their difficulties, or outlining their goals for the future is one less minute available for practising long division, diagramming sentences, memorising the periodic table, or learning C++. Recent decades have seen "back to basics" movements come and go (Weiss, 2005; Schoenfeld 1992, p. 336), and political movements aimed at eradicating unnecessary "social and emotional" learning from mathematics textbooks (Gross, 2022). At the same time, there have also been movements that recognized the potential of "learning to learn" (e.g., Goodburn et al., 2005) to benefit rather than detract from domain-based achievement, by encouraging and enabling students to take ownership of and responsibility for their own learning. Indeed, higher-level and strategic thinking may be as important for adapting to future work and life challenges as subject-based knowledge and skills. Inviting pupils to become active participants in structuring their learning also sends the message that school is *for* them as much as for the good of society.

This thesis brings together evidence on an area of educational theory, research, and practice that has grown enormously during my own lifetime, and therefore has personal significance: metacognition (MC) and self-regulated learning (SRL). At its core, metacognition emphasises awareness and control of one's own thinking, involving evaluation of oneself and one's abilities within a specific context or task. Self-regulated learning, on the other hand, highlights strategic adaptivity in learning and performance situations, often relying on motivation and self-efficacy. Yet there are many areas of overlap between these two concepts. Since around 1980, these two ideas have come to have a global impact on teaching and learning, one which has been felt through structured programs of research, as well as through informal teaching approaches and school and district level pushes to boost "thinking skills" (Higgins et al., 2004) along with measurable, subject-based learning. There have been studies that strongly supported such efforts, as well as those with less impressive results. Aggregations of research evidence including systematic reviews, meta-analyses, and overviews are valuable tools that have demonstrated the ability of MC/SRL to benefit learning in many contexts.

However, there have also been factors that have limited the uptake or effectiveness of MC/SRL approaches. In addition to concerns about test scores and "back to basics" movements, theoretical misalignments and ambiguities, competing models and terminology, and questions about assessing MC/SRL have also made it difficult to capitalise on the presumed value of these theories for school-based learning. Multiple "branded" approaches to teaching MC/SRL exist, and many approaches require substantial time, training, and flexibility to fully implement in the classroom. Before considering the evidence for the effect of these two theories on mathematics learning in primary and secondary schools, it is worth asking the question: *Why should we expect theories of metacognition and self-regulation to make a difference to learning?*

Simply put, MC/SRL offers the potential to re-write our approach to learning tasks, the goals toward which they are aimed, and the balance of power and responsibility in the classroom. Not only this, but MC/SRL promises to aid students in tackling unpredicted obstacles for which school cannot directly prepare them, and to carve out life-paths characterised by reflectiveness, intentionality, and efficacy. No subject-based knowledge or skill claims so much. The perceived centrality of these MC/SRL theories has arisen over time, and research evidence has confirmed many of these promises. Still, there have been variable effects of incorporating MC/SRL theories in the classroom, and this variability may relate to key points that have been arrived at through empirical research and theoretical development, discussed below.

In the next section of this chapter, a background of the development of theories of metacognition and self-regulated learning is presented, with evidence from laboratory and school-based research. I explain why I consider MC/SRL a single, overarching construct in this review. Next, the special relevance of MC/SRL for mathematics learning is discussed, along with potential costs and challenges of implementing them in school. Chapter 2 presents findings from previous syntheses, including systematic reviews and meta-analyses, addressing the effectiveness of MC/SRL approaches for raising achievement. In light of these, the rationale for the current study is outlined. In Chapter 3, the research questions and methods for the systematic review are described. I present the review results in Chapter 4, first focusing on qualitative aspects of the included studies and MC/SRL programmes and then the combined effects on mathematics and potential moderators of effect. Finally, Chapter 5 discusses the results, limitations, and potential applications of the current study findings, as well as recommendations for future empirical and theoretical work.

1.2 Development of meta-cognitive and self-regulated learning theories

This section considers the development of metacognition and self-regulated learning as objects of theory, research, and classroom practice. Attention is paid to the particular role of MC/SRL within mathematics learning, and a specific rationale for MC/SRL interventions in school-age mathematics is developed.

1.2.1 Early psychology, behaviourism, and constructivism

> With the advent of democracy and modern industrial conditions, it is impossible to foretell definitely just what civilization will be twenty years from now. Hence it is impossible to prepare the child for any precise set of conditions. To prepare him for the future life means to give him command of himself; it means so to train him that he will have the full and ready use of all his capacities; that his eye and ear and hand may be tools ready to command, that his judgment may be capable of grasping the conditions under which it has to work, and the executive forces be trained to act economically and efficiently. (Dewey, 2019, My pedagogic creed, p. 37)

To understand the importance of metacognition as impacting achievement through regulating thought, thought itself must be recognized as a central determiner of action. Ideas of self-directed learning and goal-seeking are woven throughout “classic” literature in the areas of philosophy, religion, and politics, among others, and this classical basis for reflection and self-directed learning is still referred to by education researchers today (e.g., Bond & Ellis, 2013; Baliram & Ellis, 2019). The extended quote above from Dewey, originally published in 1897, illustrates his aim to make the burgeoning field of general schooling one based on self-control and judgement rather than academic knowledge and skills *per se*. Though Dewey was a philosophical naturalist (Bishop, 2021) and considered the state and education more important than religion for improving individuals and society (Dewey, 2019, Moral principles in education, p. 3), Judeo-Christian ideals of thinking and good character still echo in his writing. In these systems of thought, the reality of the mind and its malleability were unquestioned, and to become truly educated was to become, first of all, a better thinker. In terms of school pedagogy, Dewey recommended teaching led by children's natural interests and developmental affordances, framing subject-based knowledge within a social and practical context, rather than following a rigid curriculum of book-learning (Dewey, 2019, My pedagogic creed, p. 42-44). Still, Dewey's educational writings concerned a general philosophy and approach, and he did not investigate what might be effective for daily teaching.

Psychology at this time was dominated by structuralism and functionalism as competing schools (Cherry, 2021). Structuralism sought to categorise the basic elements of thinking through the first psychological experiments, in which participants trained in introspective techniques responded to stimuli (Lopez-Garrido, 2021). Inspired by Darwin's naturalism, functionalists like Dewey were more concerned with the roles fulfilled at an individual and social level by the "coordination" of the senses and action (Dewey, 1896, p. 370). Importantly for educational theory, while structuralism emphasised universals, functionalism had more concern for differences between individuals (Cherry, 2021). Despite their differences, both schools of psychology relied mainly on subjective methods and suffered from a lack of acceptance as legitimate science (Jastrow, 1929).

Behaviourists, next to emerge on the scene, called for external manipulation and measurement to make psychological explanations more reliable and practically useful. Just as physics, biology, and chemistry exploited nature to produce desired structures, lifeforms, and reactions, psychological experiments with animals and humans were undertaken to produce desirable behaviours. Control, rather than explanation, was paramount. Hard-line behaviourists, such as John Watson (1913), eschewed the concept of mind as making any notable contribution to this project:

> Psychology as the behaviorist views it is a purely objective experimental branch of natural science. Its theoretical goal is the prediction and control of behavior. Introspection forms no essential part of its methods, nor is the scientific value of its data

> dependent upon the readiness with which they lend themselves to interpretation in terms of consciousness. (p. 158)

Behavioural methods soon came to dominate psychology, even if allusions to consciousness never fully disappeared. Notably, Watson downplayed affection and traditional nurturing as important for child-rearing. According to Watson, avoiding emotional coddling freed children up to do their work of interacting with and learning about the natural world systematically (Houk, 2000). In fact, he believed a cold and logical parenting approach would be more conducive to children's emotional and mental balance (Bigelow & Morris, 2001, p. 27). He also believed preferable behaviours could be produced in children through routine and reinforcement without using overtly moral terms.

Watson's advice, based on his own parenting techniques and experiments in infant and child conditioning (Houk, 2000), was taken up by a new generation of parents looking to make their domestic life more scientific (Bigelow & Morris, 2001, p. 26). Such ideas also had implications for learning in school. From a behaviourist perspective, learning is merely the training of associations (Tomic, 1993, p. 38), and there is no fundamental difference in this process between a bird or rat and a child, or between children of different ages. Through experimentation, it should be possible to uncover optimal methods of reinforcement to train students efficiently in any desired curriculum or to prepare for any profession, without reference to their subjective interests or motivations. Watson complained that it was futile to wait, as recommended by Dewey, for "hidden possibilities of unfolding" (Watson & Watson, 1928, p. 40) to be revealed in a child: "The behaviorists believe that there is nothing from within to develop" (p. 41). At the same time, Watson assigns some importance to a child's understanding of the world, for example when he recommends frequent parental dialogues to uncover the child's current "verbal consolidations" (Watson & Watson, 1928, p. 160) about natural processes like reproduction. With a focus on how students actually respond to teaching, behaviourism made some lasting contributions to education, such as the "mastery learning" and "behaviour modification" approaches (Tomic, 1993, p. 43). Still, behaviourism was unable to account for differential learning capacities in different species, such as a human's ability to learn natural language (p. 44). At a more fundamental level, it is undeniable that the concept of mind and the communication of mental and emotional experiences has been important throughout history, and behaviourism has little to explain these.

Although it is commonly believed that behaviourism has now undergone a complete rejection, Roediger (2004) points out that behaviourist influence on psychology remains in how it relies on experimental verification of theories. Piaget, according to Schoenfeld (1992, p. 346) was largely responsible for the rehabilitation of the concept of mind in psychology, especially its development over time. In the interpretation of Von Glasersfeld (1982), Piaget's mental schema are based on testing and finding ideas that work, more or less, to maintain one's equilibrium within a certain context (p. 619). This could be seen as internalisation of the behaviourist approach, but it originates from the frame of reference of the thinker, rather than appealing to an external and measurable reality. Indeed, in Von Glaserfeld's interpretation of Piaget, one's recognition of and even communication with others is nothing more than a judgement about the continued "viability" of such schema (p. 625).

Despite his "genetic epistemology" leaving little route for verification of "facts" as such, Piaget was captivated by observable and consistent patterns of behaviour and understanding in children of different ages. His categorisation of the stages of human mental development, and the conditions beneficial for growth, have greatly influenced contemporary pedagogical approaches. In Piaget's presentation of learning, children's representations of the world become increasingly complex and abstract as they age. This is not a completely automatic process, however, as it depends on their encountering specific phenomena sufficient to disrupt their current representations and spur them to adapt to reach equilibrium again (Bandura, 1991, p. 258). Conscious deliberation on the meaning of experiences is also key, and it is through this interactive process that knowledge is constructed. In later accounts of constructivism, emphasis is placed on the need for schools to be nurturing environments where children are supported to develop new knowledge through cognitive conflict and shared discussions with peers, rather than passively accepting such knowledge from an authority (Hendry, 1996). Key in such an environment is teachers' ability to ". . . generate inferences and form judgements about children's developing knowledge . . ." (Hendry, 1996, p. 33), in order to devise activities to spur future growth. They must also provide opportunities for students to make their thinking explicit and reflect on its suitability for themselves (Hendry, 1996, p. 33).

Constructivism discouraged students from taking a subservient position in the classroom and rebranded their misconceptions as understandings that had some, albeit limited, usefulness. With its focus on "viable" rather than "correct" knowledge, there was the sense external assessment should be downplayed and that even less able students could offer valuable contributions in the classroom, such as giving feedback on others' ideas (Hendry, 1996, p. 32). However, the relativism seemingly inherent in constructivist pedagogy, with all knowledge seen as tentative and with no truly objective standard, could call into question the specific knowledge goals learners should aim to reach. Nor is it the case that constructivism has been able to completely quash the performative and competitive tendencies of schooling, even if it has come to represent the pedagogical ideal. During the same period that constructivism was growing in influence, for example, ways of measuring and ranking mental abilities were also being taken up. Since the late 19th century, access to educational, vocational, and life opportunities has been controlled through the use of standardised intelligence and achievement tests (White & Hall, 1980). This points to a belief that the kinds of mental development society values are not only linked to either training, as in behaviourism, or opportunities and encouragement to actively reconstitute understanding, as in constructivism. The belief is that they are constrained by the nature of the mind/brain itself, with individual differences in genotype and phenotype seen as largely predictive of outcomes. Even those concerned with providing more opportunities for those considered less capable essentially demonstrate this logic when they seek strategies for improving mental functioning.

1.2.2 “Meta” concepts in psychology and pedagogy

It was against this backdrop that the first “meta” in psychology arose, metamemory (Brown, 1977, p. 2), which is concerned with learners’ adeptness with memory strategies and their understanding of their own storage and retrieval of information. Flavell, Friedrichs, and Hoyt (1970) studied normally-achieving schoolchildren and found age-related differences in memory performance and predictive accuracy. On the other hand, Brown (1977; 1980), who initially participated in behaviourist, animal-learning research (cf. Brown, 1994), went on to work with mainly learning-disabled children, prompting them to remember lists of items, numbers, pictures, or elements of a story. Children’s own, un-prompted strategies for remembering were also considered, and optimal strategies were elicited or taught and assessed over time. In her work, Brown (1977) indicates a strong link between memory abilities and development, stating that at certain stages of maturation, strategies can be acquired untaught, while the use of trained strategies will fade over time if learners are not developmentally ready to maintain them.

Research on metamemory from this period highlights the general failure of young participants to learn, apply, maintain, transfer, or generalise memory strategies. Frequently children are presented as ignorant or even dishonest (Brown, 1977, p. 65) about their own cognitive awareness and strategies, showing a clear deficit. Yet it must be remembered that the laboratory studies and even the structured learning environment of school may not represent optimal cognitive contexts for very young children. Subpar performance in memory tasks could be a symptom of poor ability, or it might be due to a lack of interest in the decontextualised tasks, a point Brown (1977; 1994) herself makes. She also shows a concern that this line of investigation could decrease opportunities for disabled students. Rather than debate the possible “structural” (p. 67) impairments, researchers should work toward effective interventions that teach generally applicable cognitive skills like “self-interrogation” (p. 97), which could include checking one’s knowledge of a task, searching for more information if needed, and evaluating performance. In fact, Brown comes close to equating such skills, when used spontaneously and appropriately for a given task, with intelligence itself.

Building on metamemory research, Flavell (1979) presented metacognition as a “*new* [emphasis added] area of cognitive-developmental inquiry” even though the term was in use some years prior (e.g., Brown, 1977, p. 4). While metamemory was more limited in scope, metacognition signified higher-level knowledge and control of all types of mental processes. Flavell (1979), who was strongly influenced by Piaget, sought to formalise children’s metacognition into a model that could explain and predict differential functioning across and within individuals. He was also influenced by the concept of theory of mind (Flavell, 2000), which refers to the beliefs an individual has about their own and others’ mental qualities, formed through observation and self-reflection. Research on babies and young children show that theories of mind become increasingly sophisticated as they age. It has even been suggested that primates display a simple theory of mind (Flavell, 2000).

Despite the basic simplicity of metacognition as a concept, Flavell (1979) drew early attention to the complex web of interactions it represented. According to Flavell (1979), metacognition can be categorised into knowledge and experiences, cognitive tasks, and the actions or strategies used to accomplish such tasks. Consider, for example, his detailed categorisation within just the category of metacognitive knowledge, shown in Table 1. Although other aspects of his “model” are not specified in as much detail at this point, Flavell (1979) makes clear his goal is to demonstrate the interdependence of the components, as well as their practical usefulness. His focus is on metacognition, not simply for its own sake, but in the service of a more mature management of thinking and behaviour, and this is a key theme that is repeated throughout the theoretical literature on metacognition. It also pre-figures later efforts to connect MC and SRL.

Category	Sub-components	Explanation
Person	Intraindividual differences	Knowledge or beliefs about how one thinks or learns in different contexts
	Interindividual differences	Knowledge or beliefs about how others learn or think, as different from oneself
	Universals of cognition	Knowledge or beliefs about commonalities of thinking and learning
Task	Available information	Knowledge or beliefs about task-related information and how it influences one's task approach
	Demands or goals	Knowledge or beliefs about task criteria and difficulty
Strategy	Cognitive	Knowledge or beliefs about strategies for making cognitive progress
	Metacognitive	Knowledge or beliefs about monitoring cognitive progress

Table 1. Illustration of metacognitive knowledge from Flavell's 1979 model of metacognition, including categories, subcomponents, explanations, and examples.

Other researchers would criticise Flavell's early version of MC as too unspecified and all-encompassing--Schoenfeld (1992, p. 347) calls it a "kitchen-sink definition"--yet from the beginning, Flavell was clear that metacognition involved two central components: knowledge or information about thinking and active processes to control thinking. These two essential aspects of metacognition would go on to contribute to expansive models of self-regulated learning that also include the learner's self-system (e.g., motivation, self-efficacy, and affect), and the social setting and culture in which learning takes place. While each of these plays a role and will be discussed below, it should be emphasised that accurate and relevant metacognitive knowledge is at the core of the SRL system. In fact, "metacognitive knowledge" was added as a new category to Krathwohl's (2002) revision of Bloom's taxonomy, showing the broad reach of this concept into educational theory[1].

Even with its inherent complexity and potential to encompass many elements of a learning situation or task, metacognition is still distinct from other related theories that arose around the same time. For example, theory of mind involves knowing general characteristics of minds, but metacognition is concerned with how specific minds function within the context of

[1] Krathwohl (2002, p. 214) does not cite Flavell, instead referring to concurrent work by Pintrich, but he does utilise Flavell's sub-categories of person- ("self"), task-, and strategy-related metacognitive knowledge.

“tasks” or problems. Metacognitive knowledge is separated by Flavell (1979) into knowledge about people (self, others, humans generally), about tasks (their nature, requirements, and difficulty), and about strategies or actions (cognitive and metacognitive). This knowledge, according to Flavell, can sometimes be accessed without one's awareness within routine activities, which would align with Winne's (1995) "inherent features of SRL” or "automatic inferences,” but it can also be brought to the forefront of the mind during metacognitive experiences, which is the next category Flavell (1979) discusses.

Metacognitive experiences are most likely to occur adjacent to or during a specific task in which there is a moderate, but not overwhelming, level of challenge and something of value at stake. Intuitions about one's progress or likely success in a task are examples of metacognitive experiences that can draw on, as well as update, our metacognitive knowledge. Strategies they trigger can be aimed at considering the problem, ourselves as problem-solvers, or both. Although Flavell (1979) includes tasks and strategies in the same list as metacognitive knowledge and metacognitive experiences, he does not cover them extensively except as subcategories of the other two. Task and strategy knowledge is key to self-regulatory actions, and are covered in more detail below, but whether such knowledge is always *metacognitive* is not clear. Likewise, metacognitive knowledge is not always accurate, according to Garofalo and Lester (1985), nor is it always helpful. In the context of developing metacognitive knowledge to improve learning, Pintrich (2002) sees it as essential that students develop a language to discuss metacognition with teachers and peers, which they can also use in their own self-reflections and monitoring/evaluation of cognitive functioning.

Flavell’s (1979) original model of metacognition needed clarification and demonstration in research and practice. One issue is the potential overlap between Flavell’s different categories of metacognition (i.e., knowledge, experiences, tasks, and strategies). Garofalo (1986) notes that an item of metacognitive knowledge may combine information about specific tasks and strategies, for example. Flavell (1979) himself asserts that any of the four aspects may be cued successively or simultaneously in a real-life situation. He also points out that the same cognitive or metacognitive move may serve multiple aims, or it may not serve the desired aims at all. Monitoring may be necessary but not sufficient to benefit learning or task performance, yet Flavell (1979) argues that increased monitoring is more likely than the alternative to lead to constructive aims and " . . . wise and thoughtful life decisions . . ." (p. 910). Flavell’s (1979) model was also limited in its support from empirical research and teaching practice. At the time, there had been few attempts to test the accuracy or predictive power of learners’ metacognition within learning settings, and Flavell (1979) does not offer clear pedagogical recommendations for building students' metacognitive abilities. However, he does suggest a possible developmental sequence of metacognitive skills which could be looked for in the classroom. According to Flavell (1979), a child's general awareness of "knowing" would precede an awareness of different kinds and levels of knowledge, which would be followed by an awareness that knowledge may be false, incomplete, or difficult to acquire. Flavell states that the child's initially limited MC abilities should be seen as "building blocks" (p. 909) for the future, rather than as simply inadequate. Kuhn (2000) agrees with the idea of a developmental progression of metacognition, and she even argues that metacognitive awareness enables or constrains other kinds of cognitive growth, as learners suppress outmoded task approaches and reinforce newly acquired ones (cf. Siegler, 1994).

Although detailed classroom interventions based on MC/SRL had not been tried when Flavell (1979) published his model, Brown (1980) reports on metacognitive aspects of reading comprehension, such as checking if a text agrees with one's current knowledge or underlining important passages. Brown (1980) notes that mature readers seemed to use such techniques spontaneously, while others did not apply reading strategies even when trained. Greater cognitive development seemed linked to more effective reading behaviours, but Brown (1980) admits some of the reported tasks might not resemble authentic school reading activities. Several years later, Palinscar and Brown (1984) developed one of the earliest MC/SRL-based reading interventions, Reciprocal Teaching (RT), which would later be expanded into Fostering Communities of Learners (FCL). Both RT and FCL utilised a Vygotskyan approach, in which a less able student engages in dialogue with a relative "expert" to internalise productive learning strategies and improve performance in school. Palinscar and Brown (1984) emphasise that the RT intervention is designed to probe students' "zones of proximal development" (Vygotsky, 1978, cited in Palinscar and Brown, 1984), where learners cannot yet perform tasks alone but can with sufficient scaffolding. As competence grows, the scaffolding must be judiciously decreased so there is always an optimum level of challenge and a chance to test their skills. Although commonly used now, these ideas were just becoming influential in education and applied psychology at the time of writing.

One key point of Palinscar and Brown (1984) is that it describes an increasingly realistic application of the RT intervention, first with a pilot stage that is reported only briefly, then with a researcher-led small group intervention compared to an alternate intervention and two non-synchronous control groups, and finally a teacher-led intervention with intact groups of remedial or low-ability readers. Within the RT intervention, teachers and students took turns completing certain comprehension tasks related to a set passage of text, such as summarising, asking "teacher-like" questions, clarifying, and predicting. It is notable that RT did not directly ask students to reflect on their thinking patterns *per se*, or to consider how they would use these comprehension strategies in other work. Still, the RT treatment was associated with impressive gains in almost all measures relative to the alternative intervention, no treatment, and test-only comparison groups. There is also evidence that RT assisted the participants in other subjects, such as science and social studies. Although RT is named for the exchange of typical teacher-student roles, there is only clear evidence that this was happening in the final stage of the study, when RT was implemented by teachers in groups of four to seven students. In these groups, transcripts show multiple students taking the lead at different points to summarise the passage and ask questions, while other students offer suggestions and criticisms, and the teacher intervenes much less often. These patterns of interaction support the implication by Palinscar and Brown (1984) that the teacher is actively responding to the students' growing skills by removing scaffolds and increasing demands on performance, and this points to the importance of input and buy-in from practitioners when developing interventions. What is less clear is whether the positive effects of RT are attributable to the group dialogue, the reading strategies themselves, the regular feedback offered to students, or an increase in self-monitoring. Palinscar and Brown (1984) acknowledge this uncertainty and state their intentions to focus on the mechanisms of action now that the RT intervention has had some success.

While early efforts to implement metacognition in teaching were underway, the concept of metacognition (and a closely related concept, self-regulated learning) were still in need of expansion and clarification, and there were numerous efforts to do this throughout the 1980s and 1990s. Reeve and Brown (1984) present Flavell's version of metacognition as being primarily about an individual's *conscious* knowledge of thinking, though this knowledge is not necessarily always utilised effectively, while in their own "information-processing approach" (p. 346), metacognition begins as implicit and can be consciously controlled through internalisation of social interactions with more mature thinkers. For Reeve and Brown, metacognition is closely tied to executive function, and they use the terms almost interchangeably. Problem-solving is the aim of metacognition/executive function for them. Reeve and Brown (1984) lament that little research to that point had fully considered individuals' *development* of metacognition throughout life, but they reference Piaget and Vygotsky in their explanation of social-interactive development. They state that effective teaching should include learners in shaping the purposes of such interaction, should differentiate between learners of different needs, should remove scaffolds as learners grow, and should try to facilitate internalisation of scaffolds. Although Reeve and Brown acknowledge the tentative nature of their developmental theory of metacognition, they present evidence from their own experiments and those of others that interventions based on this theory can be effective in expanding students' regulation of their learning and self-concepts within a specific learning domain (i.e., reading and writing), and that this growth can be maintained over time, as shown in the RT research. In future metacognition research, Reeve and Brown state they would support a deeper consideration of learners' perceptions of competence and how these affect task performance, a theme that would be taken up by Zimmerman (1990) and other researchers under the heading of self-regulated learning, to which I turn next.

1.2.3 Self-regulated learning

It is clear that more than person, task, and strategy knowledge is necessary to be successful in school (Pintrich & De Groot, 1990, p. 33). By the late 1990s, models of metacognition (e.g., Borkowski, 1996) were expanding to include motivational and self-system management, that is, how individuals "self-regulate" within a learning situation. While according to Schoenfeld (1992), Bloom's original taxonomy (1956, cited in Schoenfeld, 1992, p. 358) reflected a period of "sharply delineated distinction" between affective and cognitive research, self-regulated learning and self-efficacy theories sought to reconcile the two strands of work. Researchers increasingly recognized that learners' goals, attributions for success or failure, and self-efficacy can have important influences on their actions (Dweck, 1986; Zimmerman, 1995; Efklides, 2009, p. 80). In fact, affective components may be key sources of information about the meaning and value of tasks, as well as about a learner's capabilities and likely success. In writing about self-efficacy, Bandura (1977; cf. Schunk, 1991) states that individuals acquire and process information from many sources, including beliefs, prior experiences, observations, and discussions, and they compare and weigh this information to make judgements about their capabilities. This again reflects the information processing approach to human cognition (Borkowski, 1996), in which external stimuli are believed to affect behaviour not directly, but through individuals' interpretations and choices based on the information they have available. Bandura (1977) also considers self-efficacy judgments an element of "social learning" because so much of the information individuals utilise originates in social contexts. These judgments result in different kinds and amounts of effort, which lead to different task performances.

Like metacognition, self-regulation may sometimes be inadequate or inconsistent (Pintrich, 2000, p. 452). Winne (1996) argues that learners are in fact always "self-regulating" in some way, even when there is little prompting to do so, but because their values and information may differ greatly from those of adult stakeholders, performance may not meet external standards. Nor do learners always meet their own performance goals (Dweck, 1986). Winne (1996) presents possible "sites" where individual differences in self-regulation can be seen. Because of the interconnectedness of the factors in his model, a difference in one factor could be amplified or compensated for by differences in the others. Like Flavell (1979), Winne (1996) begins with knowledge as a critical component, both knowledge of the general domain and task, as well as knowledge about potentially useful approaches. Winne (1996) notes that experts in a domain more than novices may accomplish many ordinary tasks with little need to overtly self-regulate; however, experts also develop an extensive repertoire of strategic patterns or “chunks” that support complex efforts to regulate in novel or challenging tasks (Perkins & Salomon, 1989; Schoenfeld, 1992; Zimmerman, 2002, p. 66). Knowledge of strategies or tactics includes not only the procedures that should be carried out, but also information about aspects of the learning situation that cue those procedures, which Winne (1996) calls “IFs” (p. 336), but this is often referred to as conditional knowledge (Moshman, 2018, p. 600). Without helpful conditional knowledge, learners may misapply or fail to apply strategies they know.

While knowledge is important in SRL models, active regulatory processes are the main focus. Any student may occasionally use techniques to boost motivation, suppress distractions, or increase effort, but Zimmerman (1990; 2002) presents a truly self-regulated student as one who employs such means systematically, consistently, and in a way that is tailored to specific tasks. Whether in metacognitive or SRL explanations, regulatory behaviours can be separated into planning, monitoring and evaluating of efforts. These are often thought of as pre-, during-, and post-performance activities, but there may not always be a linear progression between them (Pintrich, 2000, p. 455). Pre-performance activities can include task analysis, dividing the larger task into sub-tasks, establishing monitoring systems, and allocating resources. Discrete task analysis and goal-setting seem to benefit performance on shorter as well as extended tasks. Students’ conceptualizations of academic tasks can vary based on their prior experiences in the domain, level of expertise, and cultural values (Schoenfeld, 1992). Winne (1995) demonstrated that students who viewed reading and comprehension of academic texts as a speedy and straightforward process (i.e., "quick learning") were more likely to under-utilise effort, to employ simplistic strategies like memorization, and to overestimate success, while students who saw it as slow and complex better regulated their efforts to the task demands. Pre-performance regulation can also include insuring against likely problems. Lewis (1989) found that training college students to graphically represent the operations required in mathematical word problems prior to solving them avoided a common reversal error. Related to goal-setting, Bandura and Schunk (1981) found that having young children set challenging short-term rather than long-term goals for self-directed learning of arithmetic increased their performance and improved self-efficacy, motivation, and calibration. Dweck (1986) and Borkowski (1992) believe that learners with mastery rather than performance goals are more likely to have incremental ability views and higher self-efficacy, and Dweck (1986) states such learners will more readily transfer strategies to new contexts. Despite this evidence about the importance of goal-setting, learners in school contexts may have little freedom to set their own goals, which may discourage SRL efforts.

Monitoring to check progress towards a goal, also discussed in metacognitive models, is another area where individual differences in self-regulated learning can be seen. According to Winne (1996), learners differ in their propensities to monitor and evaluate. SRL actions can be effort-intensive, and learners need to perceive the likely outcomes as justifying the effort, known as “expectancy-value theory” (Pintrich, 1999, p. 467). Some learners display “perfectionist” tendencies in Winne’s (1996) view, and they are not satisfied unless they reach a high confidence in their performance, while others are more tolerant of ambiguity regarding their performance. In challenging tasks that require frequent modification of effort and strategy, individuals need to balance the resource demands of monitoring against task-completion, and there may be maturational limits to how well learners do this (Zimmerman, 1990). Several researchers have discussed whether monitoring too early with a domain or task could impede acquisition but there does not seem to be consensus on this issue. Monitoring can also be affected by aspects of the task situation itself, such as time pressure (Winne, 1996), and the amount and type of monitoring will directly influence learners’ regulatory choices. Structured self-questioning is a monitoring strategy that has been used successfully within RT (Palinscar & Brown, 1984) and other MC/SRL approaches.

Feedback from others and self-attributions are two other valuable sources of SRL input that may often come following a learning activity. Although Borkowski (1996) states that feedback can spur SRL behaviours through motivation, feedback can also inform individuals about the suitability of their approaches and level of goal-achievement. Feedback from teachers that responds to students’ current patterns of thinking rather than the correctness of answers may lead to better self-regulation and learning (Cardelle-Elawar, 1990). Not only do students' perceptions of their own learning matter, but teachers' "working models" of their students' thinking also have a large impact in the classroom, according to Borkowski (1992). Cardelle-Elawar’s (1990) work with mathematics teachers illustrates that teachers need not only domain expertise but experience in analysing and discussing models of thought within the domain, including misconceptions and ineffective strategies, in order to give productive feedback, a theme also reflected in Schoenfeld (1987; 1992). Feedback usually refers to others' assessments of students' work, but students' own assessments can provide important SRL information as well. Attributions, or explanations for success or failure, can have a strong motivational effect. Bandura (1991) states: "The effects of causal attributions on motivation and performance attainments are mediated almost entirely through changes in self-efficacy beliefs" (p. 258). Attributions based on effort or acquired skills are more likely to increase self-efficacy beliefs than those based on chance or unchangeable attributes. Bandura's (1991) social-cognitive theory proposes that self-efficacy drive is a major reason people undertake challenging endeavours, and this theory is frequently referred to in MC/SRL-based mathematics research.

1.2.4 MC/SRL as an overarching construct

From the preceding section, it is evident that metacognitive and self-regulation theories, as applied to academic learning, share common features. Metacognition highlights learners' awareness, but action, as well as reflection, is at the heart of Flavell's (1979) and Brown's (1977) views of metacognition as an essential tool for mature functioning. Self-regulation theories add an emphasis on emotions, motivations, and the social milieu of school, but self-awareness and adaptive strategies (i.e., metacognitive knowledge) are still considered key aspects of self-regulation. Both models assume that individuals do not always use meta-level skills in a manner that is consistent or appropriate to a task, but both also claim that explicit training can help such skills develop. Under both systems, training helps learners "externalise" (Brown, 1997, p. 402) mental processes and act strategically to reach goals. Additionally, authors of research reports frequently refer to both groups of theories, and interventions based on these theories often feature similar activities. For these reasons, I consider MC/SRL as a single, overarching construct within this review, and the implications of this choice are discussed further in the final chapter.

1.3 Metacognition and self-regulation in mathematics

The previous section demonstrated how theories of metacognition and self-regulation might make a difference to learning generally. It also demonstrated why I have grouped metacognitive and self-regulated learning together to form the theoretical base for the research synthesised here. Next, the role of MC/SRL in mathematics specifically can be examined from a theoretical and practical perspective. This section is not exhaustive but outlines several potential benefits from incorporating MC/SRL approaches in school mathematics, as well as some of the expected costs and potential challenges of such approaches. In the following chapter, I consider extant reviews of MC/SRL approaches that show measurable benefits on academic achievement in order to set the context for the current review.

1.3.1 Benefits of MC/SRL-based mathematics approaches

First, approaches based on MC/SRL theories might be important for developing a mature understanding of the nature of mathematics, of which Schoenfeld (1992) is a major proponent. Although mathematics achievement is highly valued within the school context and is a frequent prerequisite for upper-level STEM courses, there is still a perception that achievement results from a simple combination of ability and effort within a relatively straightforward learning sequence. This impoverished view of mathematics learning leaves little room for adaptability, creativity, or the construction of new knowledge by pupils (cf. Kajander, 1999). It also presents mathematics as a largely solitary activity. Thus, mathematics learning in school fails to reflect the approach of actual mathematicians. Schoenfeld (1992) suggests that mathematics learning be re-conceptualized from acquisition of content to development of a mathematical perspective on the world and membership within a community of practice (i.e., "enculturation," p. 340; cf. Cobb et al., 1997, p. 152), as well as the ability to be creative within the domain. Students would achieve this through grappling with extended, non-routine problems, working collaboratively, and engaging in dialogue that is not simply "Socratic" (Schoenfeld, 2020, p. 1164), that is, designed to lead to the correct answer. In short, Schoenfeld (1992) advocates aiming for students' development of an identity as mathematicians, and for classroom activities to reflect realistic mathematics activities without a predefined correct answer. This pedagogical ideal is also echoed by Davidson, Deuser, and Sternberg (1994), who state that adaptive problem-solving approaches and creating the conditions for mathematical insights should also be taught (cf. Kramarski & Mevarech, 2003, p. 302). A broader conceptualization of mathematics teaching as identity formation requires students to take a much more active role in the classroom, and it would benefit from or necessitate the metacognitive and self-regulatory techniques discussed above.

Metacognitive and SRL approaches would also be predicted to benefit learners' beliefs about the process of learning mathematics according to Garofalo (1989) and Mayer (1998). In "traditional" mathematics teaching, students are judged primarily based on their operational and problem-solving performance and less so on conceptual knowledge and beliefs. Students can display mistaken mathematical assumptions in their problem-solving approaches and persist in an unproductive approach because they believe it is the only acceptable way to solve a certain type of problem (Schoenfeld, 1992, p. 356). If they believe they have carried out the correct procedures, they may fail to check the reasonableness of their responses at a deeper level or seek empirical confirmation. Therefore, a focus on "correctness" might not always lead to the best performance. Prompting students to reflect on and explain their problem-solving efforts may lead to more correct responses and more importantly may be instrumental in convincing them they can make sense of the mathematical principles at work. Research has indicated that even mathematics teachers do not always demonstrate productive beliefs about mathematics (Garofalo, 1989, p. 502) and may resist teaching it in a more reflective way that decentres correctness and their own status as authoritative sources of knowledge. The success of MC/SRL approaches to mathematics, therefore, may depend on teachers' own reconceptualisations of the discipline and their roles as teachers, which is discussed further in the next section.

Mayer (1998) argues that three interrelated competencies are important for mathematics achievement, "skill," "metaskill," and "will" (p. 50). Relating to the third category, will, metacognition and self-regulated learning could support a more balanced affective and motivational perspective. Mathematics performance situations trigger anxiety and intimidation in many students. Students often believe mathematics is a "hard" discipline, in which activities are designed to make them struggle or to test their innate ability. Cardelle-Elewar (1995, p. 91) states that lower ability students display minimal persistence in the face of difficulty in problem-solving. This is logical from a traditional mathematics perspective in which ability is signalled by rapid insight and extra effort will not produce results. Within a MC/SRL framework, especially those focusing on "meta-affect," students could be taught strategies for recognizing and coping with feelings of difficulty and frustration when they arise. They might also be taught to retrain their attributions from ability-based to effort-based ones, to adopt an internal locus of control, and to recognize their own achievements without comparing them to others'. Students from diverse backgrounds, who may face barriers to developing their identities as competent mathematicians, might especially benefit from MC/SRL training to deal with negative affect and demotivation. For example, it seems that males are more likely to estimate their mathematics abilities highly and be less thrown-off by challenging mathematics tasks (Seegers & Boekaerts, 1996), which may lead to gender-imbalance in school mathematics achievement. MC/SRL approaches that emphasise all students can achieve in mathematics through focused and strategic efforts may help to mitigate this. Destigmatizing asking questions, producing errors, or asking for help in the mathematics classroom could also be components of an MC/SRL approach (Cardelle-Elewar, 1995, p. 93). While such an approach empowers students to take an active role in their own growth rather than simply relying on teachers to guide them, it also encourages collaboration more than competition and the opportunity to use peers as a resource (Brown, 1994). As Stanbridge (1990, cited in Hendry, 1996, p. 32) points out, even students who may not be high achievers on individual assessments can make valuable contributions to the class through voicing their struggles and challenging others' ideas. Altogether, such changes triggered through an MC/SRL approach may produce a more supportive social atmosphere and boost pupils' confidence and self-efficacy in mathematics learning.

The building and organising of a mathematics knowledge base is another area in which MC/SRL approaches could benefit students. To become truly proficient in mathematics, learners need to be able to make connections between different mathematics principles and operations, and to apply strategies whenever they might be fruitful, regardless of the original context in which they were learned. This is known as the “transfer” problem (Mayer, 1998, p. 49; Schoenfeld, 1992, p. 352, Desoete, Roeyers, & De Clercq, 2003, p. 197), and it has been lamented that individuals frequently fail to apply knowledge across domains even when it would be appropriate and helpful. This could be exacerbated by the extent to which higher-level training, particularly in mathematics, becomes increasingly specialised and disconnected from what has come before. Pupils may believe there are separate rules for every type of problem presented to them (Garofalo, 1989, p. 503), and fail to learn from analogies, as recommended by Pólya (1954, p. 13)[2]. As discussed above, it would be better to convince students of logic and interrelatedness of mathematical systems and principles, so that they persist in trying to “connect the dots.” In a MC/SRL-based approach, students would be prompted to explicitly consider ways in which new knowledge is aligned with or diverges from what they currently accept to be true. They would also be led to work through inconsistencies in their mathematical beliefs, and to build up potentially weaker areas of their knowledge, exposed through dialogue and attempts to practically apply it.

Finally, MC/SRL approaches would be expected to assist pupils in a core mathematics learning activity, solving routine and non-routine problems (Mayer 1998, p. 49). Problem-solving, as Schoenfeld (1992, p. 337) laments, is an underspecified concept, but one that is the heart of school mathematics learning. He outlines several common rationales for problem-

[2] Pólya (1954, p. 13) states, “. . . two *systems* are analogous, if they *agree in clearly definable relations of their respective parts*” (emphasis original). He argues for the usefulness of examining simpler systems, figures, or equations when trying to make sense of more complex ones. Yet choosing appropriate analogies is constrained by knowledge and intuition developed through experience in the domain, so his recommendations may not be as helpful for novice mathematicians and students still building their knowledge base.

solving, including as a type of “recreation” (p. 338) or to practise recently taught principles and strategies. According to him, most problems are posed inauthentically, leading students to ignore the context, or “window dressing” (p. 342) and focus on the operations alone, especially those taught most recently. Most school mathematics problems have no real-world importance and have a single correct answer, which the teacher already knows. Pupils also believe they should be able to solve them quickly by executing the appropriate operations (Garofalo & Lester, 1985, p. 167), without the need for deep consideration, an impression sometimes reinforced by teachers themselves (Schoenfeld, 1992, p. 359). Even in traditional problem-solving, raising MC awareness can help students “manage” and “coordinate” (Garofalo & Lester, 1985, p. 169) the complex operations necessary, as discussed below. However, MC/SRL approaches reach beyond computation to prepare students for real-world problems, in which judgments must be made as to what relevance mathematics has, whether the operations required are simple or complex, and the need to be strategic in other, non-quantitative ways. Whereas in school mathematics, solving an equation may be the end goal of a problem, in life this is often just the start of the solution. In MC/SRL-based mathematics, students would be much more frequently exposed to real problems without a pre-set answer, requiring ingenuity, creativity, patience, and collaboration. They would be exposed to problems they do not yet have the competence to solve, or for which there is no precise solution, and they would be encouraged to set their own problems.

Not only do MC/SRL approaches seek to provide a richer problem-solving experience for students, in which problems are used to build more than “display” knowledge, they may also train strategies that raise performance on traditional assessments. One such strategy is to check understanding of a problem before attempting a solution. Students need to be prepared to deal with both well- and ill-posed problems (see Table 2), and to consider the context carefully rather than casting it aside. Mathematical language and complex problem-statements may be especially challenging for learners with special needs or second language learners, and Cardelle-Elewar (1990), citing a problem-solving model by Mayer (pp. 166-167), trained teachers to support the latter type of student through metacognitive instruction and feedback. Students were led to identify whether their errors were related to misinterpretation of the problem or difficulties with calculation, indicating the appropriate corrective actions. Another approach, IMPROVE (Mevarech & Kramarski, 1997), has also been used to help pupils make sense of and address mathematical tasks. In IMPROVE, pupils are prompted to restate a problem or use a graph to represent the mathematical situation and the information provided, to relate the problem to others they have encountered, and to evaluate exactly what they are being asked to do. Only then should pupils choose from among possible strategies and plan their problem-solving approach. In such approaches, students can be led to estimate a reasonable solution and use this to check results. MC/SRL-based strategies can also be taught for monitoring and "control" during problem-solving, though Schoenfeld (1992) points out a lack of theoretical and research-based agreement on what effective control looks like (p. 364). Building students' task- and person-related MC knowledge could help them predict when errors are more likely and when monitoring efforts are necessary (Garofalo & Lester, 1985, p. 168), such as in extended tasks requiring multiple calculations.

Type of problem	**Problem statement**	**Underlying operation(s)**	**Solution strategy**
Well-posed	George and Khaled are baking cupcakes for a 200-person charity event. They want to be sure every person can buy 2 cupcakes, if desired, and each of them will bake half the cupcakes. George takes 2 hours to bake 100 cupcakes, while Khaled takes 3 hours. How much time will be spent in total on baking?	200 x 2 = 400 total cupcakes, or 200 cupcakes for each baker. (200 / 100) x 2 = 4 hours for George. (200 / 100) x 3 = 6 hours for Khaled. 4 + 6 = 10 hours total.	Ignore the context and isolate and complete the required computations.
Ill-posed	George and Khaled need to bring 400 cupcakes to a charity event, and they agree to each bake half. George wants to save time by using boxed cake mix and premade frosting, but his cupcakes may not be as popular, and they would cost 40% more to produce. Last year, 300 made-from-scratch cupcakes sold out within 3 hours of the 5-hour event. This year the event will be 4 hours long. Khaled refuses to use pre-made ingredients. What should George do?	400 / 2 = 200 cupcakes per baker. George's profit = (cupcakes sold X price) - (cupcakes sold x cost x 1.4). 300 / 3 = 100 scratch-made cupcakes sold per hour. (Several more operations are possible, using estimated quantities).	Consider the context and compare last year's and this year's events. Determine the rate of sale for scratch-made cupcakes and estimate their profitability. Estimate the rate of sale for George's cupcakes and calculate their relative profitability. Argue for a course of action.

Table 2. Examples of well- and ill-posed problems.

1.3.2 Costs and potential challenges of MC/SRL-based mathematics approaches

Zimmerman (1986, p. 311) posits that self-regulation does not emerge spontaneously but is essentially a "culturally transmitted method for optimizing and controlling learning events," and he labels authority figures as important "socializing agents" in this process. Therefore, considerable attention must be given to training and supporting classroom teachers when adopting an MC/SRL approach. For teachers as well as students, MC/SRL may operate as a "threshold concept" (Meyer & Land, 2003), transforming their perspectives on the classroom and their role in it. Supporting MC/SRL could also be potentially "troublesome" (Meyer & Land, 2003, p. 5) because teachers may need to reconsider the nature of the mathematics discipline, how people learn it, and their own roles in the classroom. Research has indicated teachers' underlying beliefs may be open to change, but efforts may depend on how traditional their own mathematics experiences have been (Schoenfeld, 1992, p. 360). Hendry (1996, p. 32) indicates a high degree of teacher confidence is needed to avoid overly "directive" pedagogy, in which teachers present mathematics principles, engage students in problem-solving and practice, and use tests to check learning. In an MC/SRL approach, teachers need to operate at a meta- as well as practical level and see their roles as "facilitators" (Baumfield, 2006, p. 188) rather than simply instructors. They must regularly gauge pupils' mental states and assist them in confronting negative affect. They must also be open to reorganising plans in light of what students are ready to learn. Acting as models of self-questioning (Cardelle-Elawar, 1995, p. 86; cf. Callahan & Garofalo, 1987, p. 23) and decentring their own expertise could be especially challenging for teachers. In Cobb et al. (1997, pp. 165-6), a teacher in an interactive mathematics approach used her expertise to recast students' explanations of problem-solving into more standard forms, yet she made it clear they were the experts when it came to their own thinking processes. Finally, of course, teachers must spend time building pupils' metacognitive knowledge, vocabulary, and strategies. This last requirement may be the primary focus of many MC/SRL training programmes, but it is not clear how effective this would be without an accompanying shift in pedagogical philosophy and a more flexible learning environment in which students' choices really matter. Regardless, the role of the teacher in MC/SRL training is paramount, and this is covered in the narrative synthesis for the current review.

Investing in an MC/SRL programme further requires consideration of the monetary costs, which can include the potentially extensive training for teachers described above. In Cardelle-Elawar (1995, p. 84) mathematics teachers completed 17 hours of training to implement a MC/SRL approach, while in Desoete, Roeyers, & De Clercq (2003, p. 192) they completed 10 hours. Additionally, some studies have included handbooks, lesson plans, and ongoing teacher supervision, which can also add to the cost of the MC/SRL interventions. Regarding the materials for students, various MC/SRL approaches to mathematics have included interactive devices and tutoring software, manipulatives, games, books, handouts, and whiteboards, among other materials. An MC/SRL approach is more concerned with the mindset of students and teachers than with the learning materials per se, yet a more creative and flexible interpretation of mathematics is more likely to be supported with materials that facilitate collaboration, discussion, and practical application (Garofalo, 1989, p. 504). Regarding technology investment, adaptive tutoring and problem-solving software could relieve some of the burden from teachers to explicitly teach MC/SRL strategies and give feedback (Zimmerman, 1986, p. 312). Online tools could also facilitate collaboration between pupils and allow teachers to monitor and guide productive discussions. Technology could also be useful for automatically generating and giving students opportunities to respond to feedback on their strategies or solutions to tasks. Finally, a combination of hardware and software could be used to capture elements of a learning situation and record students' reflections on what they have contributed and learned (e.g., Motteram et al., 2016). The cost of such technologies is likely to be the biggest financial barrier to implementing a comprehensive MC/SRL-based approach to mathematics, but where students already have access to computers, tablets, or mobile devices, there are many affordances for implementing at least some MC/SRL strategies. As with any technology-reliant approach, planning, training, and ongoing support must also be included as investments, or teachers may revert to established practices.

Other possible challenges of implementing an MC/SRL mathematics approach could relate to students themselves. Norman (2020) states that, in some cases, being overtly metacognitive could be detrimental to students' emotional well-being and even their task performance, and that these risks are overlooked in "normative" (p. 2) research about metacognition. To begin with, students may be ill-prepared by prior school experiences to start taking responsibility for their own learning. They may be comfortable being told what they need to learn, given structured opportunities to practise, and tested in predictable and straightforward ways. Pupils may not be prepared to struggle through open-ended tasks or collaborate with peers of different ability levels, and they may shy away from revealing areas of confusion to teachers and peers. Although MC/SRL approaches, specifically those dealing with meta-affect and motivation, offer strategies for dealing with emotional states that could hamper learning, the potential for MC/SRL approaches to disrupt the established social structure and students' perceived roles should not be minimised. Teachers should therefore expect a certain degree of resistance, and they should also be explicit about the social norms and value inherent in the new approach, illustrated effectively in Cobb et al. (1997). Furthermore, since a foundation of MC/SRL approaches is a higher-level awareness and adaptivity, pupils should be oriented to the potential benefits of the new approach and led to consider contexts in which MC/SRL strategies may not be the most efficient. Structured monitoring during problem-solving, for example, could increase cognitive load and might be less effective than relying on automatic processes to complete simple operations. Nor is it always helpful to try out every possible solution to a problem. To be truly MC/SRL proficient in mathematics, learners must have a repertoire of strategies and a solid base of mathematical knowledge (Schoenfeld, 1992, pp. 351-2) at their disposal. As Norman (2020) points out, individuals must judge when it beneficial to employ MC/SRL strategies, that is, when the returns will be worth the efforts and potential discomfort.

Finally, key stakeholders, such as school administrators and parents, must be prepared to accept that MC/SRL approaches may look very different from traditional mathematics teaching, and they may not lead to higher achievement on traditional measures in the short term. MC/SRL takes time and resources to implement, and students must begin taking ownership of their learning through making real choices. They need to be allowed to experiment, question, and make mistakes without fear of repercussions or of being categorised as less mathematically able. In the same way, assessments should reflect the range of knowledge and skills developed within an MC/SRL-based approach. Until assessments look for a mature mathematics mindset as well as operational proficiency, MC/SRL approaches may or may not make a substantial difference in any particular school or classroom, although the evidence points to the fact that they often do. This is explored in the next section, as evidence for effectiveness within existing research syntheses is considered and the rationale for the current review is developed.

Chapter 2: Narrative overview of previous systematic reviews on metacognition and self-regulation in mathematics teaching[3]

Given the presumed benefits of MC/SRL approaches for mathematics learning described above, research studies and syntheses have attempted to uncover the effect of implementing them within school mathematics. To set a context for the current systematic review, a number of previous reviews were examined, shown in Table 3.

Author(s) and Publication Date	**Dates Searched or Included**	**Reports screened for inclusion**	**Reports retained in review**
Hattie, Biggs, & Purdie (1996)	1968-1992	1,415+	51
Higgins et al. (2004)	1984-2002	6,424	23 in-depth (191 mapping review)
Higgins et al. (2005)	1984-2002	318 (incl. 191 previous)	29
Dignath, Büttner, & Langfeldt (2008)	1992-2006	100 approx.	48
Dignath, & Büttner (2008)	1992-2006	not reported	74
Donker, et al (2014); de Boer, Donker, & van der Werf (2014)	2000-2011	1000+	58
Dent & Koenka (2016)	1986-2011	3,585	79
Ergen & Kanadli (2017)	2005-2014	115	21
de Boer et al. (2018)[4]	2000-2016	4,251	36
Lee et al. (2018)	1998-2017	121	18
Perry, Lundie, & Golder (2019)	2000-2017	Not reported	51 "core studies," (plus other "relevant texts")
Verschaffel, Depaepe, & Mevarech (2019)	1997-2019	109	22
Wang & Sperling (2020)	1992-2019	341	36

Table 3. Reviews included in narrative overview.

[3] A version of this section was presented at the British Education Research Association (BERA) 2021 annual conference and can be accessed at https://doi.org/10.13140/RG.2.2.15807.59047/1

[4] This review is not presented below in detail, as it deals with longer-term effects of interventions, which are not a major focus of this review.

2.1 Overview approach

Although detailed and useful for setting the context for the current review, this is not a comprehensive overview of the MC/SRL research area, for several reasons. First, the reviews included here were not gleaned through a systematic search and screening process. Instead, online searches as well as personal contacts were used to locate reviews that have had an important impact in the field. There could be other relevant reviews that were not identified. All the included reviews were completed since the mid-1990s, and they all documented the processes for finding and summarising relevant studies, but most were missing elements now accepted as important for systematic reviews. Overall, the estimated effect of MC/SRL programmes on academic outcomes has been substantial, with an effect size of around 0.6. Three of the included reviews focused on mathematics outcomes only, Lee et al. (2018), Verschaffel, Depaepe, & Mevarech (2019), and Wang and Sperling (2020). The latter two reviews only report study-level effects without combining them in a meta-analysis. Earlier reviews included effects in mathematics as well as other domains. Although Hattie et al. (1996) does not present effect sizes or summary estimates specifically in mathematics, but rather combines outcomes in all subjects, it was included because it is a landmark study that is often referred to in the other reviews.

When considering individual studies, reviews have highlighted the fact that not all metacognitive or self-regulatory interventions have led to greater learning gains, and reviews have sought to differentiate studies with regard to the types of students served, the domain specificity and theoretical basis of the intervention, the “dose” or amount of the intervention given, and other pedagogical choices such as group or individual work and the use of technology. Studies have also been examined for whether a researcher or classroom teacher has delivered the intervention and whether the study used purpose-built or standardised assessments. Comparing study designs and interventions in this way would be expected to assist practitioners in choosing the most appropriate types of metacognitive or self-regulated learning interventions for their own teaching situations, however different reviews have not shown strong agreement in the estimated effects of specific features of the interventions or study designs. It is not clear whether these differences might arise due to different search strategies, inclusion criteria, or screening procedures for studies reviewed, or whether there have been real shifts in intervention effects over time that are reflected in the reviews. Because the relevant reviews only partially overlap in terms of their conceptual frameworks and review methods, and because they cover different publication periods, the differences in review outcomes should not be very surprising. These differences will be discussed with the intention of providing a rationale for the current review to cover potential gaps in understanding around metacognition and self-regulated learning interventions in mathematics.

2.2 Overview results

The primary studies included in the reviews span the years 1968 to 2019 (see Figure 3), with 18 to 79 included reports per review. Because some reports did not make it clear which studies were included in the systematic review or meta-analysis, it is not possible to determine how much overlap there is between reviews. Even where reviews covered the same publication years, differences in the search and screen process could mean a lack of overlap. Still, some primary studies have been included in multiple reviews, and thus performing a meta-synthesis would not be appropriate. It is not always possible to determine which specific references or publication types were included in each review, but Table 4 indicates the numbers of reports for each publication type. In some cases, numbers of studies, comparisons, or effect sizes are reported rather than number of reports, and it is not always possible to determine these numbers specifically within the mathematics domain. Because unpublished research and “grey” literature is more likely to show null or negative intervention results (Song, Hooper, & Loke, 2013), a lack of balance in publication types could increase the risk of bias in the summary estimates of effect. Reviews that included mostly or all journal articles might show a higher overall effect from MC/SRL interventions. Only reviews of MC/SRL interventions with a meta-analysis for academic outcomes are included in Table 4, but all studies are discussed below.

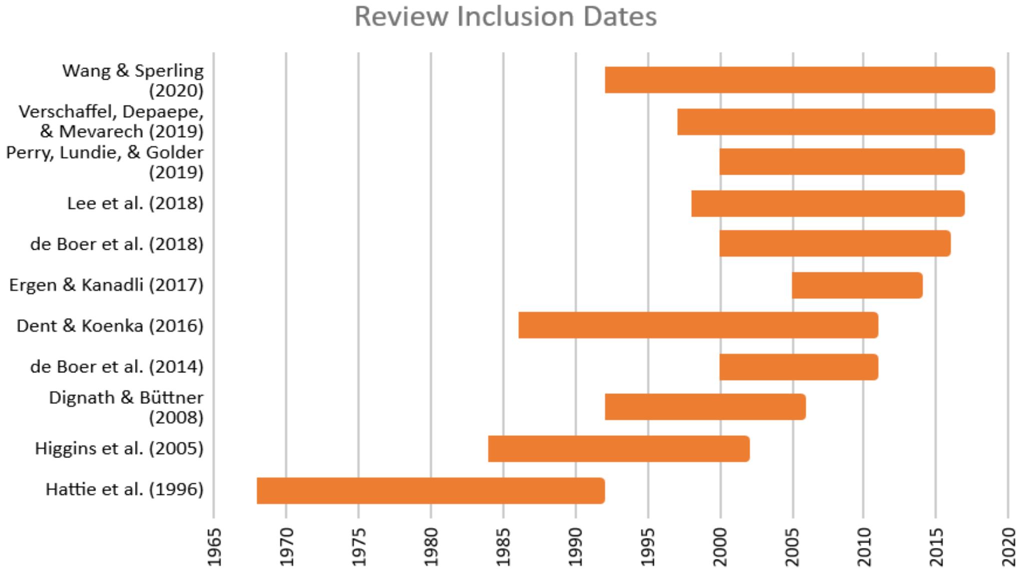


Figure 3. Included publication years of reviews in narrative overview.

Review	Total reports	Mathematics studies or effect sizes	Journal articles	Book chapters	Dissertations and theses	Conference papers	Combined effect (all subjects)	Combined effect mathematics
Hattie, Biggs, & Purdie (1996)	51	U	189 or 207[5] (effect sizes)	22 (effect sizes)	41 (effect sizes)	U	0.57 ("performance")	NR
Higgins et al. (2005)	29	9	27	2	0	0	0.62	0.89
Dignath, Büttner, & Langfeldt (2008)	30[6]	25	28[7]	0	0	1	0.62	1.00
Dignath & Büttner (2008)	46[8]	25 (primary), 12 (secondary)	46	0	0	0	0.61 (primary), 0.54 (secondary)	0.96 (primary), 0.23 (secondary)
Donker, et al. (2014); de Boer, Donker, & van der Werf, (2014)	58	44	57	0	0	1	0.66	0.66
de Boer et al. (2018)	36	8	36	0	0	0	0.12 (immediate to delayed)	0.22 (immediate to delayed)
Ergen & Kanadli (2017)	21	8	14	0	6	1	0.86	1.10
Lee et al. (2018)	18	18 (22 effect sizes)	18	0	0	0	NA	0.97

Table 4. Reviews of MC/SRL with report types and combined effects in all academic subjects and in mathematics. Reviews without a meta-analysis are not included. U=unclear from report, NR=not reported, NA=not applicable.

[5] This is reported differently on pp. 113 and 118.

[6] The authors state (p. 111) there were 48 total "studies," but this likely means comparisons.

[7] Numbers are based on only 29 asterisked items in the review reference list.

[8] On p. 243, the authors state there were 74 "studies," but this likely means comparisons.

2.2.1 Hattie, Biggs, and Purdie (1996)

Like the current review, Hattie, Biggs, & Purdie (1996) sought to compare and combine effects from a "disparate" group of "study skills interventions" (p. 99). They state: "These interventions have aimed at enhancing motivation, mnemonic skills, self-regulation, study-related skills such as time management, and even general ability itself; creating positive attitudes toward both content and context; and minimising learning pathologies" (pp. 99-100). Thus, MC/SRL-type interventions would clearly be included. By the mid-1990s, interest in "learning to learn" and broadly metacognitive approaches had gained momentum in educational practice, to the extent that Hattie, Biggs, and Purdie (1996) collected 51 intervention studies for their review and meta-analysis. Their review has become a commonly referred to benchmark for individual studies and subsequent reviews in this research area. Although not labelled as a systematic review, Hattie et al. (1996) reported the general inclusion criteria, the databases and dates searched, and the keywords for their review but did not report on the screening process and how many studies were excluded with reasons at each stage. There were then fewer commonly accepted guidelines for performing a systematic review, and this term was often conflated with "meta-analysis" as it seems to be here. This review is still systematic, however, since it uses pre-specified criteria and search strategies. In addition, Hattie et al. (1996) scanned reference lists of included reports for studies missed by the initial search. This "snowballing" may increase the review coverage but is less replicable and could introduce bias. Only two databases were searched, ERIC (Educational Resources Information Center) and Psychological Abstracts. While several inclusions are theses or dissertations, the great majority are academic journal articles. The dates range from 1968 to 1992, although the authors state that they searched the databases for reports published between 1983 and 1992. Earlier studies may have been identified through snowballing. The included studies in Hattie et al. (1996) needed to provide information sufficient to calculate an effect size, operationalised as between group differences or within-group changes in performance.

Rather than classifying interventions based on theoretical foundation, Hattie, Biggs & Purdie (1996) chose to use the SOLO taxonomy, which categorises interventions based on their "structural elements," being either "prestructural," "unistructural," "multistructural," "relational," or "extended abstract."[9] They also reported whether each intervention aimed for near or far transfer of skills. Transfer refers to the application of skills to a new context that is conceptually

[9] The authors note they did not use the "prestructural" category, as such interventions would be automatically excluded as "unsatisfactory" (p. 104).

similar or distinct from the original one, and this can be achieved with either a “high” or “low road” (Perkins & Salomon, 1989). The SOLO classifiers are explained in detail by the authors and relate mainly to the complexity in and integrative nature of the strategies to be learned. As with transfer, the SOLO taxonomy has not been employed the later reviews covered here, and it is doubtful that practitioners would be assisted by these categories to choose an intervention. Hattie, Biggs, and Purdie (1996) also note whether the outcome of interest is “reproductive” (e.g., memorising domain content) or “transformational” and measures “performance,” “memory,” “attitude,” or “study skills.” They also note the “thrust” of the intervention, described as either “structural aids,” “memorization,” “study skills,” “motivation,” or a “Feuerstein” program (i.e., instrumental enrichment). Although the authors explain these categories, their boundaries and applications to included studies have some ambiguity. All included studies are classified as self- or teacher-directed, implying there were no researcher-led programmes. The academic domain of each study is not reported.

As mentioned, effect sizes were calculated for all included studies. Aggregate outcomes are reported by the number of effect sizes (n=270 total), not the number of studies, and these could represent multiple assessments or multiple treatment groups. In some cases, the total effect sizes within a category (e.g., “publication form,” “test quality,” “design of study,” p. 113) do not sum to 270. Predictably, “study skills” (n=106) accounted for the most effect sizes under “program thrust,” and “extended abstract” (n=40) was the SOLO classification with the most effect sizes. The other SOLO categories were applied to between 16 and 29 effect sizes, showing diversity in the interventions. “Performance” (n=157) was the most common outcome, with most outcomes being “transformational” (n=122) rather than “reproductive” (n=92), or “other” (n=56). Notably, most effects were associated with students of high ability (n=109), while low ability and underachieving students were the next common. As distinct from most other reviews, studies at university level are included in Hattie, Biggs, and Purdie (1996), and they account for more effect sizes (n=103) than those at each of the lower grade-levels. The authors infer that university participants chose their participation status (p. 112), which could impact the effect size interpretation. Overwhelmingly, effects were from teacher-directed interventions (n=204) more than self-directed ones (n=29). Most assessments were considered high quality, with seven effects from low quality tests being excluded from further analysis (p. 112).

Combined effects are reported for all outcomes (ES=0.45) and with the study as the analysis unit (ES=0.63). Effects did not differ substantially based on whether there was a comparison group (ES=0.42) or whether the intervention group’s own pretest was used to calculate effect size (ES=0.48), but designs labelled “other” showed somewhat higher effects

(ES=0.59, p. 116). Publication type also showed a correlation, with journal articles having the highest effect (ES=0.55), the next most effective being books/book chapters (ES=0.41), and dissertations (ES=0.0) being “not effective” (pp. 116-118). Within the studies remaining after excluding those with low quality tests, study quality in general had no impact on effect (p. 118). Due to having multiple effect sizes for each study, it is sometimes difficult to interpret the variation in effects. For example, under program “thrust,” attribution is reported as accounting for 11 effect sizes and having a higher-than-average combined effect (ES=1.05, p. 118), yet elsewhere it is shown these all relate to the same primary study (p. 111). The combined effects also do not distinguish between academic domains, although the reviewers do report effect size differences for performance (ES=0.57), affective (ES=0.48), and study skills related assessments (ES=0.17). Related to performance, “extended abstract” or “Feuerstein” programmes were found to be particularly effective on measures like Raven’s Progressive Matrices (p. 121), and this could have a large influence on the combined effects due to the higher number of included effects in these categories (p. 113). Still, the authors report that “unistructural” interventions are the most effective (ES=0.84) for performance outcomes possibly due to their focus on simple, immediately applicable study skills (p. 116). Because “performance” here may include language arts, science, history, and even general cognitive skills in addition to mathematics, the effects reported in Hattie, Biggs, and Purdie (1996) are not directly comparable to those of the current review.

2.2.2 Higgins et al. (2004, 2005)

The EPPI-Centre in the UK released two technical reports describing a systematic review and meta-analysis of “thinking skills” programmes, which “require learners to articulate and evaluate specific learning approaches; and/or…identify specific cognitive and related affective or conative processes that are amenable to instruction” (Higgins et al., 2004, p. v). As in the Hattie, Biggs, and Purdie (1996) review, Higgins et al. (2004) point out the difficulties of precisely defining “thinking skills,” but their focus is on giving practitioners evidence with which to choose among programmes, especially commercially available ones (p. 8). Thus, rather than a detailed theoretical model, Higgins et al. (2004) present five categories of such interventions, each epitomised by a well-known MC/SRL-type programme being used in schools. For example, under “cognitive operations,” they discuss Feuerstein’s Instrumental Enrichment, while under “heuristics (strategies),” they refer to de Bono’s Cognitive Research Trust programmes. The other categories comprise “formal thinking,” “thinking as manipulation of language and symbols,” and “thinking about thinking: metacognition” (p. 9). However, the reviewers state the challenge of neatly categorising interventions has only increased as interventions “infuse” and combine approaches (Higgins et al., 2004, p. 10).

To execute their review, Higgins et al. (2004) report a comprehensive electronic search process, in addition to direct contact and reference list-searching, which resulted in 8,053 initial returns to be screened for inclusion. After screening 681 full-text reports, Higgins et al. (2004) identified 191 reports for their “systematic map” (p. 20), which outlined key aspects of thinking skills studies, such as the educational contexts, samples, and subject areas. They give details of each study’s approach and outcomes in the appendix. Over half of reports originated in the US (34%) or the UK (27%), but all reports needed to be in English (p. 21). Importantly, studies could involve a variety of settings and subject areas, including arts, physical education, and religion, but they needed to focus on the years of compulsory schooling (ages 5-16). More reports described secondary (45%) than primary contexts (34%, p. 22), but early secondary (ages 11-13) was the most common. Both pupils and teachers could be the focus, although it was more often the former (p. 24), and the design did not need to be an intentional manipulation of the thinking skills programme with a control group (p. 71). In fact, Higgins et al. (2004) indicate that 40% of included reports were of “naturally occurring intervention[s]” (p. 29), although most (n=150 reports) involved some kind of quantitative data collection, mainly attainment scores, with only 41 reports being purely qualitative. Forty-eight percent of reports involved researcher-led evaluation (p. 28). Half (51%) used the regular teacher to implement the programme, while researchers implemented it 14% of the time, and 25% of reports did not specify implementer. As with the current review, this is a concerning omission, since teachers’ roles in supporting the new skills during normal class sessions could influence outcomes. Higgins et al. (2004, p. 40) state teachers’ shaping of classroom discussions of thinking is likely to be important. Half (49%) of reports discussed a named programme, such as Philosophy for Children, Cognitive Acceleration through Science Education (CASE), or Feuerstein’s Instrumental Enrichment (p. 25), but the remainder of reports did not feature a pre-packaged intervention. Thinking skills were taught as part of a regular class (i.e., “infused") most of the time (n=111), rather than being given in separate sessions (i.e “enrichment"), but this was not coded for all included reports (p. 25). Science (34%), literacy (20%), and mathematics (19%) were the most common subjects, with all others accounting for 6% or fewer reports each (p. 23). The substantial focus on science makes sense, given that CASE or its mathematics equivalent, CAME, accounted for more reports (13%, p. 25) than any other “branded” intervention.

To carry out an “in-depth review” (Higgins et al., 2004) and meta-analysis (Higgins et al., 2005), the researchers whittled down the initial set of 191 reports, first by considering only those with both qualitative and quantitative outcomes and a “researcher manipulated intervention” (2004, p. 31). This left 23 studies, whose characteristics generally reflected those of the larger group, except there is a higher proportion of primary level studies (p. 32) and those using named programmes. Each study is presented with a rating in several quality-related categories, though the reviewers state that underreporting, especially of qualitative details, was a limiting factor along with actual errors, and these contributed to seven studies being judged as “low weight of evidence” (p. 38). Half of the studies were judged as “medium,” and the rest “high weight of evidence” (p. 38). Higgins et al. (2004) presents preliminary findings about trends in effectiveness based on the in-depth review, but Higgins et al. (2005) includes a full meta-analysis. Here, 29 studies are included, since the original systematic searches were updated and the inclusion criteria shifted slightly, such as requiring data sufficient to generate effect sizes and at least 10 student participants (p. 12). The combined effect from such studies is reported as ES=0.62 (p. 28) for both cognitive and curricular outcomes, whereas for affective outcomes it is much higher at ES=1.44, but this was with a subgroup of only six studies. In mathematics, the combined effect is ES=0.89 from nine studies, which is higher than in science (ES=0.78) and almost twice that of outcomes in reading (ES=0.49, p. 32), but the mathematics effects were much more variable than the other two. Although a range of interventions were included, the reviewers found that those training “metacognitive strategies” were associated with higher effects (ES=0.96) than either Instrumental Enrichment (ES=0.58) or Cognitive Acceleration (ES=0.61) programmes. These trends support the focus of the current review on MC/SRL-based interventions in mathematics.

Higgins et al. (2004) were able to double-rate a portion of the studies to strengthen confidence in their results, and they report “a high level of reliability” (p. 17). In some cases, differences arose when one team member coded a higher level of detail than another (p. 29) or there was uncertainty in coding distinctions. Higgins et al. (2004, p. 29) state that the 19 keywords under “thinking skills” were the most challenging to apply consistently across the research team. Raters were instructed to select up to three codes from a list that includes, for example, “critical thinking, “higher order thinking,” “logical thinking,” and “systems thinking” (p. 72), which could be a challenge to differentiate, or, if referring to the authors’ own labels, could vary in their applications from study to study. Even though this could limit the interpretation of their review, the intervention labels chosen would likely be recognized by practitioners and other stakeholders, more than would the SOLO taxonomy labels utilised in Hattie, Biggs, and Purdie (1996). Overall, Higgins et al. (2004, 2005) is a user-friendly review; the search and screening process is comprehensive and transparent, and the qualitative and quantitative reporting balances judiciousness and helpful detail. The reviewers demonstrate awareness of the drawbacks of some of their choices, and their claims about effectiveness are balanced. Still, their findings related to mathematics are based on a small subset of studies, with unclear comparability to the current review.

2.2.3 Dignath, Büttner, & Langfeldt (2008); Dignath & Büttner (2008)

Coming several years after Higgins et al. (2005), Dignath, Büttner, & Langfeldt, (2008) is the first review of self-regulated learning (SRL) interventions with mathematics outcomes, though it is limited to programmes for primary level learners. Being more clearly defined than either "study skills," or "thinking skills," SRL involves ". . .cognitive, metacognitive and motivational processes, which work together during information processing" (p. 104). Learners' goals and adaptation to the learning context are also important. References are made to works by Winne, Pintrich, Zimmerman, and Bandura, among others, showing that Dignath, Büttner, & Langfeldt, (2008) rely on these different researchers' common focus on domain-level and meta-level functioning, and on the regulation of learners' affect. These emphases are still comparable to much more recent reviews, covered below. Unlike some other reviews, Dignath, Büttner, & Langfeldt, (2008) consider metacognition an aspect of self-regulation. In fact, they posit three levels of functioning for the learner: the cognitive as the lowest "information processing" level, the metacognitive directing the cognitive, and the motivational at the highest level, providing the impetus for strategic regulation (Boekaerts, 1999, cited in Dignath, Büttner, & Langfeldt, 2008, p. 104). These assumptions generally align with the current review, although I am more concerned with areas of overlap between MC/SRL-type models than their distinctions.

To execute their review, Dignath, Büttner, & Langfeldt, (2008) used a systematic search of both English- and German-language academic databases. Their numerous keywords included "study skills," "learning to learn," and "thinking skills" (p. 104), indicating high overlap with the central constructs of Hattie, Biggs, and Purdie (1996) and Higgins et al. (2004, 2005). In fact, the authors were concerned that self-regulated learning alone was too "fuzzy" (p. 104) a term to capture all relevant studies, but checking alignment with SRL was part of the process for screening "about 100 articles" (p. 105) from the searches. Additionally, studies had to involve mixed or average ability classes within a normal school setting led by teachers or researchers, so laboratory studies, those using computer-based teaching (CT), and those involving students with special designations were excluded (p. 105). English-, German-, or French-language reports were admissible. Control groups were also required, which is a more stringent design requirement than Hattie, Biggs, & Purdie (1996). The included reports were dated between 1992 and 2006, and the authors report them with their theoretical bases, treatments, academic subjects, number of sessions and outcomes (pp. 123-126). The main focus of the report is the meta-analysis and moderator analysis. The authors included multiple effect sizes per article, leading to 263 total effect sizes, but they combined those for the same construct within one article and adjusted sample sizes when the same control group was used for multiple comparisons (p. 109). Three main categories dealt with academic outcomes: mathematics, language arts, and other subjects. The authors also considered cognitive/metacognitive and motivational outcomes. Dignath, Büttner, & Langfeldt, (2008) did not differentiate cognitive versus metacognitive strategies within outcome assessment as they stated it would be especially challenging to assess when students were being metacognitive (p. 106), but they did distinguish these strategies with the SRL interventions. Motivational strategies were also considered as part of the SRL training, and subcategories under the three main strategy categories were also coded. Cognitive strategies could involve repetition, elaboration, organisation, or problem solving; metacognition could involve knowledge and skills, like planning and checking; and motivation could involve making attributions, feedback, and "action control" (pp. 107-108). The intention was to use these categories and other aspects of the study as moderators to home in on the potential reasons for variations in effects.

The combined effect for all outcomes was ES=0.69, with a combined effect for academic performance of ES=0.62. The review included 25 total effect sizes for mathematics performance from nine total studies, with a combined effect of ES=1.00 (95% CI=0.75 to 1.24). Notably, not all studies taking place within the mathematics classroom had a mathematics performance assessment. The combined effect in mathematics performance was larger than in any other category for this review, but there was greater uncertainty in this effect given it is based on fewer primary studies. In terms of other potential moderators, there was no clear impact on all outcomes based on the type of instructed strategy, intervention length, or pupil year level, nor is there a clear pattern of effectiveness for researcher versus teacher implementation in mathematics, though scores in other measures were higher when the instruction was led by researchers. Regarding theoretical background, the review found lower effects for those programmes based on motivational rather than metacognitive or social-cognitive theories, and this was particularly true for mathematics performance and motivational outcomes (p. 113). The latter finding is unexpected. Effect sizes based on strategies and combinations of strategies are also reported within each domain area, and those including cognitive strategies were associated with the worst outcomes in mathematics (p. 113). In most outcomes including mathematics, programmes without group work performed better than those with it. Although the reviewers draw implications about what effective SRL programmes “should” include (p. 121), such as specific strategy types, it needs to be remembered that some analyses were done with small numbers of primary studies per group. As with Hattie, Biggs, and Purdie (1996), using multiple analyses increases the risk of spurious findings. Another limitation on this review’s usefulness is that no examples from the primary studies are included to illustrate the coding categories. This would have been especially warranted when discussing the specific SRL strategies trained. Finally, the choice to exclude ICT-based programmes and those for students with exceptionalities may have made the included studies more comparable but most likely limited the number of studies to be included in each analysis. It also limits comparisons with the current review.

A second article by Dignath & Büttner (2008) expands the review to include secondary school contexts until ages 15 or 16 (p. 239). Many but not all of the original studies were included in the second review, and some new primary-level studies were added, leading to 49 total primary-level studies and 35 studies at secondary level.[10] Still, effects reported for the primary level are very similar to those of the first report. Dignath and Büttner (2008, p. 244) report a combined effect of ES=0.69 on all outcomes, using a random effects (RE) model. The SRL interventions showed a higher effect at secondary level for reading and writing performance and for strategy use. For all other outcomes, there were greater effects at primary level. In mathematics, primary level effects (ES=0.96, from 25 effect sizes) were much higher than secondary level (ES= 0.23, from 12 effect sizes), which is of importance for the current review. Other academic outcomes were considerably higher in primary (ES=0.64, from 22 original effect sizes) than secondary (ES=0.05, from 6 original effect sizes). Dignath and Büttner (2008) exclude outliers more than two standard deviations from the mean,[11] but they do not specify which effects from which studies were excluded.

To examine potential moderators of effect, Dignath and Büttner (2008) utilised a method of moderator analysis different from the previous study. Rather than synthesising effects within each potential moderator category separately (i.e., running multiple meta-analyses with different numbers of included studies and outcomes), as had been done in Dignath, Büttner, and Langfeldt (2008), here a meta-regression is used (p. 242). Moderators considered are similar to the earlier review, namely: theoretical basis (metacognitive, social-cognitive, motivational), strategies trained (metacognitive, cognitive, motivational), inclusion of metacognitive reflection, academic subject, teacher- vs. researcher-implementation of the intervention, length (number of

[10] Based on the report, it is not possible to determine which specific studies were changed.

[11] In the previous review, they had used a three-sigma cut-off.

sessions) of training, and the use of group work. Final models for each outcome only included moderators whose confidence interval did not include zero (p. 243), and only those relevant to the current review are given here. More variance was explained in the secondary-level model (85%) than in the primary-level (29%), possibly due to a higher number of included effects in the latter. Longer interventions were generally more effective, unlike in the earlier review, but there is no suggested minimum or maximum length. At primary-level, social-cognitive theories were associated with better (*B*=0.33) and motivational theories with worse (*B*=-0.38) outcomes, but at secondary level, social-cognitive theories led to worse outcomes (*B*=-1.41), as did motivational theories (*B*=-0.97). Metacognitive strategies were a positive moderator at primary-level (*B*=0.39) but a negative one at secondary-level (*B*=-0.64, p. 246). For general academic performance, cognitive strategies were not an important moderator, but in mathematics outcomes were different. At primary level, cognitive strategies and longer training had a more positive connection with mathematics performance than did metacognitive reflection (*B*=-1.08, p. 247). In contrast, for secondary-level mathematics performance (p. 248), motivational theories (*B*=0.55) had a more positive association than did metacognitive theories (reference category). Dignath and Büttner (2008) found that students' strategy-use following the included interventions was more positively associated with mathematics than with other subject areas both at primary-level and especially at secondary-level (p. 249). Group work at primary level was unimportant as a moderator and at secondary-level was a negative moderator (*B*=-0.65, p. 248). Interventions led by researchers had higher effects.

These results indicate that SRL interventions may catalyse students' strategy-use in mathematics more than other subjects. The results also imply that mathematics teachers should receive thorough training, and they should teach cognitive strategies at primary-level and motivation at secondary-level. De-emphasising metacognitive strategies diverges from the implications of the prior review. Yet several caveats should be made. First, only a small number of effects in mathematics at secondary-level were synthesised, and the model explained 94% of the variation, but the primary-level mathematics model only explained 44% (p. 247). Second, theoretical bases and taught strategies were not compared consistently in the primary studies. One study may have included metacognitive and cognitive strategies in the same treatment group, while another used them in different groups. Performance outcomes could be affected by the specific combinations of moderators, the discreteness or domain-embeddedness of the SRL approaches, or the fidelity of treatment implementation, but these cannot be judged from Dignath and Büttner (2008). It is also not clear whether metacognitive theoretical basis, strategy training, or reflection looked similar in mathematics versus other domain areas. The authors note there were discrepancies between stated theoretical bases and strategies instructed (p. 255). For example, all studies regardless of theory, had nearly equivalent use of metacognitive strategies. Social-cognitive based interventions focused on “feedback and resource management strategies” (p. 255), while motivational-based training omitted feedback and metacognitive reflection for the most part. Interventions based on metacognitive theory included cognitive and metacognitive strategies as well as reflection and problem-solving approaches, but these complex interventions may have been difficult for less mature learners to benefit from, according to Dignath and Büttner (2008). Finally, as with their earlier review, categories are explained but not illustrated with examples from the primary studies, complicating the interpretation of the findings.

2.2.4 Donker et al. (2014); de Boer, Donker, & van der Werf (2014); de Boer et al. (2018)

The next team to synthesise SRL-programme studies was Donker and colleagues, publishing two reports in 2014. A further paper in 2018 reports on the delayed effects of “metacognitive strategy instruction.” The first review considers the role of the SRL strategies themselves and the second focuses on other aspects of the research that could impact outcomes. Donker et al. quote Pintrich’s definition of self-regulated learning as “an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate and control their cognition, motivation and behaviour, guided and constrained by their goals and the contextual features in the environment’’ (Pintrich, 2000, p. 453). Thus, SRL here includes metacognitive, cognitive, motivational, and behavioural strategies, the last being an important expansion on the construct as used by Dignath and colleagues. However, Donker et al. (2014) agree that SRL is the more general construct, with metacognition being an important component of SRL. Donker et al. (2014) searched titles of English language, peer-reviewed publications in ERIC and PsycInfo. They chose 2000 as their starting date for inclusions based on foundational work by Boekaerts, Pintrich, and Zeidner being released that year (cited in Donker et al., 2014, p. 5). The search terms were simple, including only “metacognit*” and “self-reg*”. Donker et al. (2014, p. 6) state they did not search by the taught strategies because of the myriad descriptors used and the concern that relevant studies could be overlooked (p. 6). However, this decision means that those studies incorporating similar strategies but not under the SRL/MC umbrella would be excluded. With an initial return of “over 1000 articles” (p. 6), Donker et al. screened them in successive stages to include only those done in a school context, having “core” academic outcomes measured, and intentionally manipulating participants’ condition. Interventions needed to be described in detail to code the taught strategies, and some were excluded for insufficient reporting. Correlational studies were excluded as not able to lead to causal inferences (p. 6). Participants could be in primary school through grade 12 and could be average students or exceptional. Donker et al. (2014) concluded that there was no noteworthy publication bias (p. 13), even though visual inspection of their funnel plot for included articles shows some groupings of effects outside of the expected range. Because non-peer-reviewed reports were excluded, the review may be of higher quality but might not represent all research done in this area.

For the data-extraction, Donker et al. (2014) adapted a coding scheme developed by Lipsey and Wilson (2000, cited in Donker et al., 2014, p. 6), and each category is illustrated with at least one example from the included studies, which could be very helpful for teachers and researchers seeking to build on this research. Under cognitive strategies, Donker et al. (2014) list rehearsal, organization, and elaboration. Elaboration, or processing content in a way that strengthens connections with previous knowledge (p. 3), seems the broadest category, and it was also the most commonly used cognitive strategy in included studies (p. 8). Metacognitive strategies included planning, monitoring, and evaluation, and the reviewers cite Schraw and Dennison (1994, cited in Donker et al., 2014, p. 3). The authors noted that metacognitive strategies can be implemented in a general or domain-specific way (p. 6), but it is not clear this distinction was reflected in the coding. Metacognitive knowledge, covering declarative, procedural, and conditional knowledge of strategies (p. 15), was coded as "personal," such as specific feedback to the learner, or "general" (p. 7). Motivational "aspects" were coded as relating to "self-efficacy," "task value," or "goal orientation" (p. 18), but motivation as a whole was the least common intervention element (p. 8). The final category, management strategies, was divided into strategies for managing effort, peers and others, and the environment in which the learning takes place (p. 3) and was also less commonly used. While motivation and behavioural regulation are often studied independently, the requirement to be under the SRL umbrella might exclude many motivational and behavioural interventions.

After adjusting for studies with multiple interventions and outcome measures and "Windsorizing" outliers (p. 7), Donker et al. (2014) reported combined effect on academic outcomes as Hedges *g*=0.66, which is similar to other reviews. Mathematics was the most common domain and reflected the overall effect (ES=0.66 from 44 interventions), with reading (ES=0.36 from 23 interventions), writing (ES=1.25 from 16 interventions), and science (ES=0.73 from 9 interventions) being less commonly investigated. While the mathematics effect is not as high as in writing or science, the higher number of included mathematics studies makes the combined effect somewhat more certain. Regarding specific strategies, "task value" (ES=1.84, from 6 interventions) and “rehearsal” (ES=1.39, from 10 interventions, p. 9) were associated with the largest effects. However, Donker et al. (2014) also investigated the strategies in regression analyses, and this presented a different picture. Strategies trained explained 36.1% of the total variance in the regression. Considered individually, “rehearsal” had a coefficient of *B*=0.42, but it fell to *B*=0.01 when part of the simultaneous regression (pp. 9-10). “Task value” on the other hand, was still considered to be a significant predictor of performance (*B*=0.94 separate, 0.81 simultaneous regression, pp. 9-10). Donker et al. (2014) caution against over-interpretation of results based few studies (p. 10), and both task value and goal orientation (ES=0.46 from 6 interventions, *B*=-0.35 separate, *B*=-0.33 simultaneous regression, pp. 9-10) estimates relied on only 6 interventions. General metacognitive knowledge seemed to be a worthwhile strategy to include from all of the analyses, and the number of studies contributing information to this conclusion adds value. Donker et al. (2014) also found a coefficient of *B*=0.20 (simultaneous regression) for purpose-built rather than standardised assessments, which indicates outcome effects could have been biassed by the nature of the assessments used. With regard to the age/level and type of students involved, Donker et al. (2014) found that SRL interventions were somewhat more beneficial for younger students and students with special needs, although these differences were not considered significant (p. 15). Other potential moderators, such as length of the intervention and who did the training, were not analysed due to inconsistent reporting (p. 17).

Strategies were also analysed for associations with mathematics performance specifically, but many correlations are based on small numbers of studies. Only those with more than 10 studies or interventions are mentioned here. Elaboration (*B*=0.21 from 18 interventions) was the only statistically significant predictor of mathematics performance, while metacognitive strategies were the most common in mathematics studies but their value for outcomes was less clear. Monitoring (*B*=0.20, SE=0.14, from 36 interventions) was the most widely used in mathematics and had the same coefficient as elaboration but with more uncertainty. Planning (*B*=0.08, SE=0.12, from 32 interventions) and evaluation (*B*=-0.03, SE=0.11, from 21 interventions) were also frequently incorporated but appeared less effective. General metacognitive knowledge (*B*=0.03, SE=0.11, from 14 interventions) had a lower association with performance in mathematics than in other domains. Summarising the mathematics-related effects, Donker et al. (2014) conclude that elaboration should be recommended widely. As stated above, elaboration is potentially a broad category of strategies and contextualising it appropriately for specific mathematics learning objectives needs to be considered. Although the problem of “self-developed” tests was mentioned earlier, in mathematics the majority (90%) of tests were purpose-built (p. 13). Such assessments tended to show higher effects than standardised tests, but according to Donker et al. (2014), these differences were not significant in mathematics and were in the opposite direction than expected (ES=0.84 "intervention independent" and 0.61 "self-developed," p. 13).

Based on the information reported by Donker et al. (2014), it is possible to determine how many and which studies were included in each analysis, which is not possible for some other reviews. However, Donker et. (2014) did not analyse the effects of the specific combinations of strategies as they were implemented, similar to Dignath's and colleagues' reviews. From a theoretical perspective, it could be expected that the studies were designed using strategies that might be mutually reinforcing and that would fit the specific learning context best. Donker et al. (2014) also state that "...performance was almost always improved by a combination of strategies" (p. 14), and that they were also not fully able to isolate single strategy effects as they were mainly used in combination (p. 17). Another important point is to clarify what the interventions are being compared to. If the normal teaching already includes some elements of MC/SRL, then a lower intervention effect could be seen. Donker et al. (2014, p. 14) raise this issue in the context of writing but seem to overlook it in mathematics.

De Boer, Donker, and van der Werf (2014) published a further analysis of the same groups of studies, estimating the effects of implementation factors on the academic outcomes of the SRL interventions. Intervention type had the strongest relationship to outcomes. Next, unstandardised assessments showed higher intervention effects, which may be due to the narrower scope of such interventions (p. 536). Subject domain was the next most influential variable, with interventions in writing, science, and mathematics being more effective than those in reading or other domains. The third most important variable was duration of the intervention, with shorter interventions being more effective (p. 527) but minimally so and without a clear reason (p. 534). Whether or not the control group believed they were in a normal class was also a "significant" predictor, with higher effects when the control group perceived the class condition to be unusual. However, the authors stated that this variable was not always reported clearly (p. 534). Interventions led by teachers or done on a computer were somewhat less effective than those led by researchers, and, after ruling out a "novelty effect" (p. 534) the authors suggest this may be due to teachers' lower enthusiasm and expertise with the intervention (p. 535). De Boer et al. (2014) stress proper training and support for teachers implementing interventions (p. 537). Using cooperation did not improve intervention results, and the authors suggest this may be due to a ceiling effect from the interventions or because group work was often not well-utilised with both group and individual-level accountability (Slavin, 1991, cited in de Boer et al., 2014, p.

535).The fact that control groups often collaborated as well made this factor more difficult to interpret (p. 535). Randomisation was considered non-significant as a predictor, but de Boer et al. (2014) coded both group- and individually-randomised studies equally. Fidelity checks and intervention session frequency and length of the intervention sessions had no impact on outcomes. Limitations of the second study are underreporting in primary studies and small group sizes within moderator analyses. The authors could have addressed this by reducing the number of variables considered, especially since there is a risk of collinearity between some of the implementation factors. For example, randomisation and externally-validated tests may both be used by researchers concerned with robust research design, and interventions of longer duration may have fewer sessions each week to compensate. This potential collinearity is not mentioned by de Boer et al. (2014). Even with these limitations, the work of Donker, de Boer, and colleagues from 2014 is the most directly comparable to the current review. It also included a more focused look at SRL in mathematics than other previous reviews.

In 2018, de Boer et al. shifted their focus to the longer-term impacts of MC/SRL interventions, with delayed assessments ranging from four to 108 weeks following the intervention. The current review includes some studies that measured longer-term effects, but only the immediate effects are meta-analysed for consistency. Thus, de Boer et al. (2018) is not directly comparable but helps to set the context for this research. To define the limits of their review, de Boer et al. (2018) considered programmes that included at least one metacognitive component (i.e., planning, monitoring, evaluation, or knowledge, p. 101), but they also coded included studies for their use of cognitive and “management” strategies, as well as “motivational aspects” like self-efficacy, task value, and goal orientation (pp. 103, 106). In this way, the conceptual system clearly reflects that of the earlier reviews by Donker, de Boer, and colleagues (2014). Emphasising metacognition as essential for SRL, de Boer et al. (2018) state: “. . . we automatically excluded studies in which strategies were only taught as a ‘trick’, and the application of the strategy was a goal in itself instead of a means to enhance learning” (p. 101). The electronic search was restricted to English-language, peer-reviewed publications from 2000-2016, and it produced 8,744 initial returns and 36 final inclusions. Studies had to have at least a three-week delay between the intervention and follow-up assessment. (p. 101).

Included primary studies are reported in de Boer et al. (2016) with their posttest to follow-up effect size change, length of intervention and follow-up period, types of students, and the specific learning strategies focused on, but without using illustrative examples from the studies. The authors found that the follow-up period usually mirrored the length of the intervention itself (p. 106). Effect sizes were calculated through comparisons with the control group at each assessment point, and the follow up effect on academic performance (ES=0.63) was larger than the immediate effect (ES=0.50). This supports the idea that MC/SRL skills generally maintain or increase their benefits for learning after the active intervention period, and this was especially evident in mathematics. There was an ES=0.22 difference in the immediate to delayed posttest effects in mathematics, which was considerably higher than other in domains (ES=-0.03 to 0.12, p. 108). However, the mathematics effect is based on only eight interventions and apparently does not adjust for potential differences in the length of the follow-up period in each domain-based subgroup. Still, the reviewers determined that length of intervention and follow-up did not make a difference to outcomes, nor did the implementer of the MC/SRL programme (p. 108). The cognitive strategy “rehearsal” was found to be an important negative predictor of effect, as was metacognitive knowledge, and students with special needs had worse outcomes than those with low SES. These latter two findings were judged unstable, however, with a sensitivity analysis (p. 110). Rehearsal strategies could induce students to process learning material in a shallower way, according to the reviewers (p. 111). No moderator analysis within mathematics effects is reported. Finally, de Boer et al. (2018) discuss some limitations, such as high intra-subgroup variation in effects and too few primary studies per subgroup, but these would apply to many reviews of educational interventions. The overall finding, that MC/SRL interventions may have a lasting positive impact, especially in mathematics, is still important for practice.

2.2.5 Dent and Koenka (2016)

The goal of Dent and Koenka (2016) is to explain the links between achievement and self-regulated learning processes, with or without active manipulation of condition. Since this review focuses on the assessment of MC/SRL, not how it is trained, it presents a unique view of the field and is presented here in detail. The authors divide self-regulated learning into cognitive versus meta-cognitive processes, and they cast a wide net in their search strategy, described in detail and yielding 3,577 returns. The core concepts were metacognition and self-regulated learning, with grades and test scores as outcomes. Although Dent and Koenka (2016) refer to Pintrich’s (2000, cited in Dent & Koenka, 2016, p. 425) SRL definition used by prior review teams, they did not use terms related to motivational or behavioural regulation in their review, but only those related to SRL, metacognition, and cognitive strategies. In one of their searches, Dent and Koenka (2016) included outcome measures, such as the Motivated Strategies for Learning Questionnaire (Pintrich and DeGroot, 1990). It is not clear whether studies were required to have measured metacognition, self-regulated learning, or motivation to qualify for inclusion, or whether they could qualify by simply naming these as theoretical bases. It is clear, however, that Dent and Koenka (2016) include both experiments/quasi-experiments and observational studies. The focus on correlations rather than effectiveness is a major difference from the present review. Dent and Koenka (2016) also used "direct-contact strategies" (p. 441) and explain that almost 10 percent of the 79 studies were located this way. To control for publication bias, grey literature such as dissertations and theses were also included, but studies had to be from an English-dominant country (p. 442) and those included were only from the US or Canada.

Dent and Koenka (2016) calculated correlations between academic achievement and the two main self-regulated learning markers, metacognitive and cognitive strategies. Under "defining metacognitive processes" (p. 447), there were 61 studies synthesised, including studies focused on task approach and planning, self-checking or monitoring during a task, adjusting the task approach, and self-evaluation following a task. In the next analysis focusing on cognitive strategies, Dent and Koenka (2016) included 57 reports, some of which overlap with the first group to form 79 total reports meta-analysed. Cognitive strategies included both "deep processing" and "surface processing" (p. 457). Among the former would be elaboration and making connections between new content and prior learning, making inferences, and identifying main ideas, and among the latter would be rote memorization and making lists. Dent and Koenka (2016) used the original authors' descriptions and categorisation of both cognitive and metacognitive strategies for some of their analyses, but they also grouped some strategies together for other analyses. They separately discuss the correlations between metacognitive or cognitive strategies and academic achievement, then they conduct numerous moderator analyses to understand the correlations between specific types of strategies within these two groups and achievement for different ages/grade-levels and academic subjects and utilising different measures for the strategies and for achievement. They considered online and offline SRL measures, including interviews, surveys or inventories, and behaviour or speech during a task. As achievement measures, they looked at scores on standardised tests, grades, programme placement, and performance on a study task. For each of these variables, Dent and Koenka (2016) described detailed hypotheses based on previous research about the potential strength and direction of the correlations they would find, and they reflected on these hypotheses.

Dent and Koenka (2016, p. 449) found that metacognitive processes in general had a 0.20 (RE) correlation with academic achievement, with strong correlations for planning behaviours (0.30 FE and 0.38 RE, p. 450). Online measures of metacognitive processes correlated much more with achievement (0.39 RE and 0.40 FE, p. 449) than did offline measures (0.15 RE and 0.17 FE, p. 449), and standardised test scores correlated more strongly with metacognitive processes than did other achievement markers, like course grades. They also found a strong correlation between metacognition and achievement in the subject area of social studies (r = 0.34 RE and 0.31 FE, p. 449), while in mathematics the correlations were weaker (r = 0.21 RE and 0.26 FE, p. 449). The authors had hypothesised that achievement in mathematics would be less likely to be improved by metacognitive processes because it may be taught in a rigid, linear, and externally structured way, without expecting students to self-regulate (p. 434). This belief represents a divergence from the basic assumptions of this thesis, that mathematics can be better mastered when students are guided to approach it creatively and reflect on and regulate their own learning (see Schoenfeld, 1992).

Several other expectations of Dent and Koenka (2016) were not supported by their results. Metacognitive processes correlated more strongly with achievement in earlier grades, although they expected (p. 435) that MC/SRL would have a greater influence on achievement as learners matured cognitively and faced more challenging tasks. Next, they expected the "self-oriented feedback loop" (p. 430) to show a stronger correlation with achievement than did planning since monitoring and control might compensate for insufficient planning. They also thought measures of monitoring and control were more likely to probe actions critical to performance, while planning-related measures might reflect how often planning is done rather than the quality of those plans (p. 431). Overall, they still thought planning activities, such as task analysis, knowledge activation, and choosing effective strategies, would have important connections to academic outcomes. In fact, their analysis showed planning (r=0.30, p. 451) to be more strongly related to achievement than either self-checking (r=0.24, p. 451) or monitoring (r=0.09, p. 451), but when combined with goal setting, planning was found to have "the weakest correlation" (p. 451) with outcomes (r=0.01, p. 451).

Regarding cognitive rather than metacognitive processes, Dent and Koenka (2016) found as expected that the latter were more highly correlated to academic achievement than the former (0.11 RE and 0.08 FE, p. 455). This finding supports the theories of Brown (1977, p. 66) and other researchers in metacognition that posit students, especially those with special educational needs, will not fully benefit from disciplinary or task-based strategies unless guided to monitor and "generalize" them (e.g., Borkowski et al., 1989). Within this line of thinking, it is not metacognition in the form of simple awareness of thinking patterns and behaviours that can improve learning, but rather metacognition that aids the learner in optimising their use of lower-level strategies. Dent and Koenka (2016) also found higher correlations for high school rather than lower-level students and for GPA rather than standardised test score as a measure of academic achievement. Mathematics as a subject focus showed a relatively low correlation with cognitive strategies (0.07 RE and 0.05 FE), but language arts showed an even lower correlation (0.06 RE and 0.00 FE, p. 455), both of which are much lower than for metacognitive strategies. In terms of similar patterns with metacognitive strategies, cognitive strategies correlated more highly with achievement when online (0.39 RE and 0.40 FE, p. 455) rather than offline (0.15 RE and 0.17 FE, p. 455) measures were used. Dent and Koenka (2016) had hypothesised that online measures looking at students' speech or behaviour during a task and judging the amount and type of strategies used would be more accurate interview or questionnaire responses, which could be affected by social desirability factors. The results support the conclusion that consistent and observable SRL strategy use during tasks may contribute to long-term achievement.

In general, the correlations found with achievement for both cognitive and metacognitive factors were low or moderate, which makes sense given that Dent and Koenka's (2016) meta-analyses included both active and passive designs. While the current review focuses on active manipulation of metacognition and self-regulation, Dent and Koenka's (2016) review may illustrate the indirect or long-term impacts of these approaches even though it does not demonstrate causality. Regarding limitations of their research, they also mention lack of clear MC/SRL sub-concepts and labels, as well as the challenge of isolating effects of specific factors. To a large extent, these are issues inherent to the field, and they will affect the current review as well. If reports were more detailed about the operationalisation of the relevant concepts for all types of education research, then practitioners as well as researchers and reviewers would gain understanding about how key educational constructs from research align with other constructs and with classroom practices.

There are some further limitations to the synthesis by Dent and Koenka (2016) which are not mentioned by the authors themselves. First, as mentioned above, some of the included studies were found through personal contacts (p. 441), which would not be a replicable search strategy and may introduce bias. Next, although the authors did mention using only reports of US or Canada-based research (p. 442) as a limitation, this could have been avoided, had the authors been willing to include sources from outside English-dominant nations. While some relevant reports from such countries would not be published in English, many would be, given the prevalence of English as an academic language. Excluding research from Israel, for example, means that much of the relevant literature on metacognition (e.g., IMPROVE studies, such as Kramarski & Mevarech, 2003) is missing from the synthesis. Last, as with other reviews, undertaking multiple moderator analyses with the same studies raises concerns. There is an increased likelihood of a type 1 error, where "significant" findings turn out to be tenuous or unreplicable. In Dent and Koenka (2016), some analyses used small and often unequal groups of studies. For example, they found a higher correlation for online rather than offline measures (p. 449), yet this is based on 67 original correlations for offline measures and only 16 correlations for online measures. Connected to this issue, the main report does not show which specific papers were used in the moderator analysis, though one could reconstruct this from the online supplementary table. Had Dent and Koenka (2016) reported overall correlations, or focused on fewer moderators, the review may have produced more unambiguous and applicable findings. As it is, practitioners could struggle to make sense of the findings, to judge their confidence, and to strategically adapt their teaching in light of them.

2.2.6 Ergen and Kanadli (2017)

No previous meta-analyses combined effects from studies of self-regulated learning performed specifically in Turkey, and Ergen and Kanadli (2017) seek to fill this gap. However, the authors do not present detailed reasoning regarding how SRL might operate uniquely in the Turkish context, but instead they refer to social-cognitive and motivational theories and international research to support SRL teaching as generally beneficial. Like other reviews, they adopt Pintrich's (2000) definition of self-regulated learning, and while SRL is the overarching construct used in Ergen and Kanadli (2017), they include metacognitive strategies as a category for coding the primary studies. The authors report a systematic search with keywords in English and Turkish, and they utilised Google Scholar, ERIC, EBSCO, and several Turkish databases. Covering only nine years, 2005-2014, the date range is narrower than in other reviews, and they report screening 115 items, with only 21 included in the meta-analysis. Meta-analysed reports are highlighted in their reference list. In terms of research design, Ergen and Kanadli (2017) included both "empirical" (i.e., experimental or quasi-experimental) and "relational" (i.e., correlational) studies, but they excluded purely qualitative studies and reports without appropriate numerical information for generating effect sizes (p. 58). They do not mention specific review guidelines, but they do use a flowchart for items considered and included or excluded from the review similar to the PRISMA standard, and they also use a forest plot to display combined effects and a funnel plot to rule out publication bias. No quality rating is reported for the included studies, and the qualitative synthesis is minimal, with no examples to represent the coding categories. In particular, it would be helpful to see examples of how they coded the four included categories of SRL strategies, namely cognitive, metacognitive, resource management, and motivational strategies. There is a brief description of each strategy type, for example: “Resource management strategies embody . . . controlling and managing one’s time and study environment, effort, peer cooperation, and help-seeking” (p. 57). Since many MC/SRL programmes combine such strategies, it would be valuable to see how the researchers distinguished between these categories and what using each type of strategy looks like in practice. Regarding the intervention studies, who implemented the SRL programmes and for how long are not detailed. Ergen and Kanadli (2017) report an “inter-coder reliability” of 100%.

The combined effect size from this review is ES=0.86 under a RE model, with 95% CI=0.64 to 1.08. Based on Cohen’s (1988) classification (cited Ergen and Kanadli, 2017, p. 62), they assert this is a “large” effect, especially compared with earlier, international reviews. Based on finding a statistically significant amount of heterogeneity, the authors examine the qualitative categories as potential moderators of effect but report that none of these are significant. Still, the largest effects are seen with “relational” studies rather than “empirical” ones, with undergraduate rather than primary or secondary students, and with metacognitive and resource management strategies. Regarding subject area, effects in mathematics (ES=1.10) were stronger than in other subjects, and this was also the largest effect for any moderator analysed in this study. This strengthens the rationale for the current review. A final point of interest in this review is that it includes a higher number of theses, six out of 21 total studies, than in previous reviews, and this proportion is comparable to the current review. Still, Ergen and Kanadli (2017) do not find a “significant” difference in effects by publication type (p. 65), but they do not report actual differences found. For all moderator analyses, there was a small number of studies in each group, and this could contribute to the lack of statistically significant findings.

2.2.7 Perry, Lundie, & Golder (2019)

With an emphasis on educational policy, especially in the UK context, Perry, Lundie, & Golder review metacognitive intervention studies from 2000 to 2017. Here, metacognition is seen as the main construct, with SRL and “thinking skills,” for example, being sub-constructs. The authors recognize the “fuzzy quality” (p. 485) of metacognition, but they state it includes “strategies that help pupils to monitor, plan, evaluate, and regulate their performance . . . [or] solve novel problems” (p. 485). Although Perry, Lundie, and Golder (2019) report systematic searches of several online databases, they do not describe the search strings or screening processes leading to the “51 core studies” included, and they mention using “additional relevant texts” for background that were not found systematically (p. 486). It is unclear which specific studies were included, but the review included primary studies and syntheses with both qualitative and quantitative methods. The authors excluded studies that were “relatively weak” or whose samples were too small for the “strength of the claims being made” (p. 486). Perry, Lundie, & Golder (2019) do not report a meta-analysis, which is appropriate given the diversity of included designs. Instead, they report quantitative outcomes from some of the included syntheses and primary studies, though it is not clear how these were selected for reporting. They cite the EEF Toolkit, Hattie’s (2016) *Visible Learning*, and Dignath, Büttner, and Langfeldt (2008) as showing medium to high effects of metacognitive programmes. As background, the authors briefly outline the development of metacognitive theories from the work of Flavell and Vygotsky, and they discuss several “branded” MC/SRL programmes with relatively high uptake, such as Cognitive Acceleration (cited on p. 486). Perry, Lundie, & Golder (2019) lament that such programmes have not had “sufficient traction in schools” (p. 486), and the failure to assimilate metacognition into British educational policy, despite evidence for its value, is a repeating theme of this paper. They also express concern that experimental studies are being used inappropriately as a basis for policy and discuss the limitations of RCTs, but this stops short of a full critique.

In discussing the effects of metacognitive programmes, Perry, Lundie, & Golder (2019) mention various factors that could moderate effects, but how these were coded from the included studies or selected for reporting is not clear. Some factors mentioned are the embeddedness or discreteness of the metacognitive training, the length of the training, and the use of group work and assessment. They also discuss how effects might differ with pupils of different ages, ethnicities, or socio-economic backgrounds. In fact, the authors state there is evidence that metacognitive training “actually reverses the gap” (p. 491) between normal students and those considered at risk (i.e., “Pupil Premium” students). They also state that motivational strategies may reinforce metacognition and lead to more positive outcomes, but that more research on this is needed (p. 492). Without making a strong causality argument, Perry, Lundie, and Golder (2019) “suggest” the evidence for metacognitive training in schools is convincing, but that it needs to be implemented with care for the needs of the context. In fact, they state policymakers have a “moral responsibility” (p. 496) based on existing evidence to support metacognitive teaching and assessment, but the authors also recommend school leaders and teachers not wait for policy shifts before building their metacognitive knowledge and practices.

2.2.8 Verschaffel, Depaepe, & Mevarech (2019)

Focusing on ICT-based metacognitive programmes with mathematics outcomes for K-12 learners, this paper is co-authored by a researcher, Zemira Mevarech, who with other colleagues has undertaken studies using the IMPROVE approach (Mevarech & Kramarski, 1997) to MC/SRL training. The review, therefore, considers especially the effects of using IMPROVE and similar training when it is delivered in a computerised environment. As with the Perry, Lundie, and Golder (2019) review, the major construct here is metacognition, which includes both knowledge and regulation of cognition (i.e., self-regulated learning), according to the authors (p. 2). To build a rationale for using metacognitive training in mathematics, the authors refer to work by Schoenfeld and others showing it has both theoretical and practical relevance. In addition, they explain how computerised environments expanded from “drill and practice” activities to adaptive tutoring, “serious games,” and collaborative platforms, and how all of these can support both mathematics learning and metacognitive development when designed with intention. Thus, their review considers all intervention studies that use ICT as a means of enhancing mathematics-related metacognition, or that “embed” (p. 3) metacognitive guidance into ICT-based mathematics learning. Their systematic search and screening process is presented in a flowchart. It considered 109 references, a relatively small number, and it resulted in 12 included studies, to which 10 were added through “backward and forward reference search[ing]” (p. 4). Conference papers were excluded, and no book chapters or theses are cited in this review. The 22 finally-included journal articles are presented in a table summarising their designs, samples, interventions, outcomes, and findings, which is helpful for potential users of the review. In terms of qualitative synthesis, the authors report findings for different school-levels: kindergarten, elementary, and secondary. Only three studies considered outcomes of ICT-programmes with the youngest pupils, and they focused on early numeracy skills and pre-, during-, and post-task metacognition, with the later training being provided within the e-learning environment for two studies and provided by the teacher in one study. Mathematics outcomes were measured but not metacognitive ones. Due to the small sample, few conclusions can be drawn, but the reviewers caution that metacognitive activities could overtax young children’s mental resources if not done carefully (p. 5). Of the remaining included studies, most were conducted with late elementary and early secondary students, from ages ten to about 15 or 16. Within an ICT environment, elementary students practised problem-solving with an AI “tutor” or else they collaborated with other students. Secondary-level studies were similar but provided students a greater range of activities and mathematics content. Comparison groups either used the same computer programme without the embedded metacognition or in some cases received more traditional mathematics teaching without ICT or metacognitive training. Again, it was rare for the latter to be supplied by the classroom teacher rather than or in addition to the computer. Some programmes adapted either the mathematics content or the metacognitive guidance based on the student's performance, and log files were used in several studies to determine how students worked with the programmes.

Although Verschaffel, Depaepe, & Mevarech (2019) do not report a meta-analysis based on their review or do a formal moderator analysis, they do discuss possible trends of effectiveness. Overall, the ICT combined with metacognition appeared beneficial for academic and metacognitive skills, but in presenting the individual studies the authors report some null or negative results without reporting actual effect sizes. This could limit application of the findings since it is difficult to compare programmes. Factors that the authors report as possibly contributing to higher effects include a more complex metacognitive programme, greater levels of student engagement, working with lower-achieving students, and embedding the metacognitive guidance in the programme rather than having teachers implement it. Verschaffel, Depaepe, & Mevarech (2019) suggest areas for future work, such as developing programmes for more diverse mathematics skills, greater use of adaptivity and fading of metacognitive supports, and designing research to directly compare computer- versus teacher-delivery of metacognitive training and to uncover the micro-processes through which students' knowledge and skills change, leading to differential outcomes. They also recommend utilising more realistic and extended contexts for the programmes and considering motivational and affective outcomes.

2.2.9 Lee et al. (2018)

This review carried out by researchers from Texas A&M University and the University of Wyoming in the US is the only previous review discussed here to focus on a specific mathematics outcome, algebraic reasoning. The reviewers first explain that algebraic reasoning builds on arithmetic skills but requires learners to generalise and be able to represent abstract relationships between variables with mathematical symbols. These more complex skills can be challenging to develop and transfer to new situations, and difficulties may delay progress in mathematical study. Citing Schoenfeld (1987), the reviewers define metacognition as: ". . . students' thought processes and beliefs that enable them to regulate their learning activities" (p. 43). Although metacognition is the central construct here, its expression closely aligns with self-regulated learning. They also state that metacognition serves students' goals, and involves on-

and offline knowledge and skills, and can be both general and domain-specific (p. 43). Training focused on motivations and self-efficacy, or ". . . the connection between prior and new knowledge and the use of appropriate problem-solving strategies" (p. 43) is expected to be helpful in building skills in this subject area.

Through a systematic electronic search and consulting reference lists for included studies, 121 potential review inclusions were screened. Only English-language, peer-reviewed journal reports of metacognitive interventions in k-12 schools were retained, and these needed to assess algebraic reasoning as an outcome. In addition, the reviewers state that articles with numeric information insufficient to compute effect sizes were excluded. Only 18 studies with 22 effect sizes were finally included. Effects with children of different year-groups were entered separately in the analysis, leading to more effects than studies. The main focus of this short review is on the meta-analysis results. Still the authors present the included primary studies with their samples, interventions, and assessment tools. No examples are given to illustrate the qualitative categories, so it is difficult to determine the difference between metacognitive "training," "instruction," and "guidance," and this limits the applicability of the review to pedagogy. Effect sizes for the primary studies are also presented. One "outlier" is excluded from further analysis, Sun-Lin & Chiou (2017), though this seems to be due to a mistake in the reviewers' effect size calculation leading to a reported effect size of ES=22.203. Without this "outlier," there is very low reported heterogeneity (I^2=0.997), and the overall effect is ES=0.973 (SE=0.196). This is somewhat higher than other syntheses have reported, indicating metacognitive interventions may be especially valuable for teaching algebraic reasoning. However, the limited focus of the assessments means those effects might not be as apparent on more general mathematics outcomes, which are the focus of the current review.

2.2.10 Wang & Sperling (2020)

This is the most recent and most detailed review that focuses on mathematics outcomes of MC/SRL programmes but does not meta-analyse primary study effect. Wang and Sperling (2020) report the search string for each electronic database and numbers of results, but some studies were identified through Google Scholar and through reference searching which is a less replicable search strategy. They included reports from 1990 to 2020 that were "quantitative or mixed methods studies" (p. 5) of pupils in school, but they excluded book reports, qualitative studies, and those without an available, English-language full-text report. They also excluded studies of learners with special needs. Based on their flowchart, Wang and Sperling (2020, p. 6) considered 341 articles and dissertations for inclusion, resulting in a final list of 36 studies, ranging from 1992 to 2019.

Although Wang and Sperling (2020) do not report a meta-analysis, the qualitative discussion of self-regulated learning–the central construct here–and its relevance for mathematics is comprehensive and balanced. In their definition of SRL, the authors demonstrate influences from key research and theory across the history of MC/SRL:

> Self-regulated learners are active agents who use a repertoire of knowledge and strategies to regulate their learning adaptively and efficiently . . . [and] examine their strengths and weaknesses against academic task standards in order to set appropriate goals, deploy strategies, adapt to varying environments, and to overcome obstacles. (p. 1)

While not explicit here, metacognition is included in this process and is featured throughout the discussion. On the other hand, this view of SRL especially highlights how pupils intentionally interact with the teaching and learning context, and even self-reflection is viewed in light of external demands. The authors propose that mathematical representation may prove more challenging for younger students, while older students may struggle more with motivation, self-efficacy, and difficult mathematics content, indicating age-tailored SRL approaches could be most beneficial (p. 2). In addition to potential age-related differences in effects, Wang and Sperling (2020) discuss how SRL training might impact students differently based on their prior knowledge and ability level, or on the specific SRL theories and strategies employed. For example, the reviewers distinguish between cognitive, metacognitive, and motivational strategies, in alignment with other reviews, and they also differentiate between SRL models by Zimmerman, Winne, and Pintrich (Wang & Sperling, 2020, p. 7). They also discuss newer models, such as the MASRL model by Efklides and colleagues, which emphasise affective components. Based on theoretical model choice, the authors expect to see variations in the intervention elements (p. 3).

Effect sizes for mathematics outcomes in Wang and Sperling (2020) range from ES=-3.51 to ES=5.99. For each included study, Wang and Sperling (2020) also report the theory, sample, design, comparison, groups assignment method, intervention strategy and length, implementation, assessment tools, and effects in mathematics and SRL skills. They explain their coding categories and use examples from the primary study to illustrate them. Most primary studies combined metacognitive strategies with either cognitive or motivational ones, with combinations tending to be more effective. Only one study, Panaoura (2012), was coded as purely cognitive, but an examination of this study shows the intervention also prompted students to self-reflect on their problem-solving approach, which seems to belong to the metacognitive category. As for the theoretical basis, most studies relied on a social-cognitive foundation, with fewer using a metacognitive model, yet the latter demonstrated higher effects. The intervention dose ranged from 1 session to a year-long programme. There was not a clear relationship between effectiveness and intervention length or researcher- versus teacher-implementation. Wang and Sperling (2020) had planned to investigate age-related differences in effects, but included studies ranged only from 4th to 9th grade, and this relationship is not discussed further. Although geographic location is not reported for each primary study, the authors mention that the greater number of studies were done outside of the US, particularly in Germany and Israel. Israeli studies revolved around the IMPROVE approach used by Mevarech, Kramarski and colleagues, with seven total studies from this group. Most studies used researcher-developed tests, but it is not clear if this would include studies where researchers adapted a previously-validated test[12]. Researcher-developed test effects tended to be higher but less consistent than those from teacher-developed tests (p. 14). Over half of the studies used some kind of randomisation, but no formal quality-rating of included studies is done. It is not clear how relative effects were determined without executing a full meta-analysis, and the variation within each “moderator” category and strength of relationship are not clear. In addition, mathematics effect sizes were missing for nearly a quarter (n=8) of studies due to inadequate reporting in the primary studies. Still, Wang and Sperling (2020) provide value to research users by highlighting “emerging patterns” (p. 12) and shifts in focus over time.

2.3 Discussion of previous review approaches and findings

This section presented a narrative overview of previous reviews of MC/SRL-type programmes or aspects and their connections with academic outcomes. In earlier reviews (Hattie, Biggs, & Purdie, 1996; Higgins et al., 2004, 2005), the MC/SRL terminology and concepts were apparently broader, but reviews from 2008 onwards have either been based on metacognition alone or a comprehensive SRL model that includes metacognitive knowledge and skills alongside cognitive and motivational strategies, sometimes also incorporating behavioural management. Appendix 1 shows the central concepts of each review along with the search terms used to illustrate the variety of approaches to search for relevant studies. Along with the search, conceptual factors have influenced the ways included studies have been coded

[12] Some tests are reported as S+R, which presumably indicates an adapted standardised test but could indicate an average of two assessments. Only one effect size for mathematics and for SRL skills is reported in this review for each primary study.

and categorised, with reviews explaining these categories more than they illustrate their use within the actual programmes. The omission of practical detail makes it challenging to interpret the real overlap between reviews with regard to included intervention types, although it is clear some reviews include the same studies. Hattie, Biggs, and Purdie (1996) used the SOLO taxonomy and other categories to distinguish between studies, and Higgins et al. (2004, 2005) used either the name or general approach of the MC/SRL programme without specifying the actual strategies or activities. Individual strategies and broader categories (e.g., cognitive, metacognitive, motivational) have not been consistent predictors of general academic outcomes, and moderator analyses within mathematics specifically are rarely reported. Combined effects in all subjects, based on immediate posttests, have ranged from ES=0.54 to ES=0.86 for all subjects and ES=0.23 to ES=1.10 for mathematics. Mathematics effects have mainly been higher than those in other domains, where these are compared, but in some reviews, this is based on few primary studies. There has been little discussion about the types of assessments used to judge effects, and whether these are comparable between studies, but some reviews have found different effects based on how the assessment was developed. Some reviews also found differences in effect based on age or schooling level, demographic factors, or ability designations. There has been no clear relationship between intervention length and effect size.

In terms of the review methods, not all reviews have been labelled systematic, but most have included accepted elements of systematic reviews, such as structured electronic searches, clear inclusion/exclusion criteria, and the use of a researcher team to apply protocols consistently. Reviews have generally restricted the language, publication type and year, and other factors, but most have been dominated by English language peer-reviewed journal articles. Some reviews used non-replicable search strategies, such as personal contact and snowball strategies, which could admit researcher bias and make it impossible to judge the precise number of items screened for inclusion in the review. Some reviews do not clearly report which studies are included in each analysis, especially after adjustments for outliers. Where reviews have reported meta-analyses, the specific quantitative methods used have been reported in detail. Authors, especially of later reviews, have shown caution in dealing with some issues, like dependent data, heterogeneity, and risk of bias, but several reviews include multiple effects from the same study or execute multiple analyses with the same primary effects. This could lead to spurious findings and could contribute to the lack of consistency between reviews. The use of various terminology to refer to the primary studies (e.g., studies, effect sizes, interventions, comparisons, reports) also obscures the analysis methods. Finally, some reviews were limited in their scope (Ergen & Kanadli, 2017; Lee et al., 2018; Verschaffel, Depaepe, & Mevarech, 2019), or did not include a meta-analysis (Perry, Lundie, & Golder, 2019; Wang & Sperling, 2020). Dent and Koenka (2016) synthesised correlations between MC/SRL measures and academic outcomes, which could reflect the extent to which changes in achievement are actually mediated by changes in MC/SRL functioning.

It was not appropriate to perform a meta-synthesis of previous reviews due to the differences in their review methods and theoretical frameworks. Another reason is the overlap in their inclusions, which make it inappropriate to combine their summary effects. Instead, the current review has been designed in light of these previous reviews, with updating and confirming the effects of MC/SRL programmes in mathematics as a major goal. The design and methods chosen for the current review are either inspired by choices of previous reviews or seek to improve on their limitations. In particular, I increase the usefulness of the review findings by considering the actual activities and materials implemented in each MC/SRL programme, and by using these to categorise included studies rather than theoretical basis or SRL subgroup (i.e., cognitive, metacognitive, motivational). These categories are illustrated with examples from the included studies. Regarding the quantitative synthesis of effects, each study is considered the unit of analysis and contributes only one effect size in each analysis. In terms of potential moderators, only a few determined by previous research and theory are investigated to avoid type-1 errors. These and other methodological choices are detailed in the next chapter and their implications are explored in the results and discussion chapters.

Chapter 3: Methodology and methods

This chapter discusses the rationale and general methodological approach I undertook for this review. First the review questions are given, then the reasons for executing a systematic review and meta-analysis are presented. Recommended practices for research synthesis are considered, and the design and execution of the review are described in detail, starting with the search and screening process, moving to the qualitative and quantitative data extraction, the calculation of the effect sizes from primary studies, and the meta-analysis. Results of the qualitative and quantitative synthesis and a comprehensive discussion of the findings are presented in the following chapters.

3.1 Research questions

Having considered the literature related to metacognition and self-regulated learning theories, specifically how they might be beneficial for school-based mathematics learning, and the evidence for effectiveness from previous systematic reviews in the area, I devised the following questions to guide the current research:

1. During the last 15 years, what has been the effect of interventions based on theories of metacognition or self-regulation on the mathematics achievement/proficiency of school-aged learners?
2. What specific factors, if any, are correlated with higher effectiveness for such interventions?

Key term	**Working definition**
Research synthesis	Any structured approach to combining qualitative and/or quantitative results from primary studies to understand trends within a research area.
Systematic review	A pre-planned, explicit method for searching, reviewing, and reporting on relevant literature to answer a specific research question or questions.
Meta-analysis	A method for comparing and combining quantitative results from primary studies to produce a summary statistic of the pooled effect.
Narrative synthesis	An interpretive method for comparing and combining studies based on their qualitative aspects.

Table 5. Key terms and working definitions related to research synthesis.

3.2 General research approach and rationale

Once the research questions for the thesis were set, it was determined that a research synthesis, consisting of a systematic review, narrative synthesis, and meta-analysis, would be the most appropriate method to address them. My working definitions of these terms are shown in Table 5, but definitions vary in the research literature. It is worth considering the rationale for utilising this approach instead of doing a new empirical study or a traditional literature review. The first point is that a research synthesis can help stakeholders make sense of the extensive evidence that exists about teaching and learning practices. Until the mid-1990's, systematic reviews were not common in education, even though some had been performed much earlier than this in the social sciences (Higgins, 2018, p. 32). At this time, prominent academics lamented that research in general was having little impact on the direction of change in education, and that instead ". . . educational practice [was] shaped by politics, marketing, fads or other considerations . . . The failure to base educational policies and practices on evidence explains the glacial pace of change in student outcomes over time" (Slavin, 2013, p. 383). This was partly blamed on a distrust of causally-focused research and an alternative view, extant in many circles, that good teaching practices are self-evident. John Hattie, a major proponent of research synthesis, argues, "There is still a philosophy that assumes teachers know how and what data to collect to best enhance learning . . . We still teach in a manner we did 150 years ago . . . " (Hattie, 2005, p. 11). Such strong statements are open to criticism, as is the belief that experiments and systematic reviews can reveal a generalisable intervention "effect" that will hold true in any teaching context. Ignorance of contextual factors would certainly be a serious limitation on the generalisability of causal research (Morrison, 2021, pp. 178-180), yet it might be equally mistaken to assume there are so few commonalities between teaching contexts as to render causally-focused designs valueless for practitioners.

My own stance is that educational practice can be enriched through multiple forms of research, such as those that emphasise its localised, social, and personal aspects, and those that focus on broader trends. Indeed, without both of these, there is likely to be little improvement in teaching or the understanding of learning processes within the classroom. This is echoed by Hattie (2005, p. 14), and by Masters (2018, para. 7) when he says: ". . . evidence-based practice depends on the integration of reliable, local, practitioner-collected evidence with evidence from systematic, external research." Because research syntheses incorporate evidence across a range of contexts, and seek to "exhaustively" (Torgerson, Hall, & Light, 2012, p. 217) identify relevant studies, they may yield more robust information about what are *generally* effective educational practices. This generalist approach appears especially appropriate for investigating MC/SRL interventions, since this is a broad category that rests on several underlying concepts and processes, namely metacognitive knowledge and control. Only a research synthesis could identify common effects across the range of MC/SRL programmes.

When the review focus is the overall effectiveness of educational programmes, as in my first research question, there are clear advantages to systematic reviews over traditional narrative or more interpretivist reviews. First, a systematic review requires early clarification of the research questions and inclusion criteria which can make the research more straightforward to execute and interpret (Torgerson, 2003, p. 26). In a “rapid review” of 17 reports of systematic reviews in healthcare, MacLure, Paudyal, & Stewart (2016) found that only one did not have a clear research question, indicating this convention is generally followed. In contrast, traditional reviews rely on the experience and judgement of the researcher to determine what is included (Hart, 2001, p. 24) and are not necessarily led by predetermined research questions. In the current review, having focused research questions made the process more manageable for me as an independent researcher, but the results might have benefitted from an even more narrowly defined approach. Next, systematic reviews aim for detailed documentation and replicability, which even if not fully achievable can aid in interpreting the results of the research (Atkinson et al., 2014). If systematic reviews are well-reported, it is evident why some studies were included and not others, and this can illustrate where the review results could be beneficial to apply. More traditional narrative reviews, which rely on the specific background and interests of the researcher(s), may not report how included works were retrieved and analysed as so may be harder to interpret. Traditional reviews may overlook less well-known and unpublished works, especially if they rely on “snowball” searching which can amplify rather than challenge accepted knowledge. This relates to the issue of potential bias (Torgerson, Hall, & Light, 2012, p. 217). No research can be done without human interpretation, but systematic reviews are more likely to implement quality-control measures like using a team of specialists, publishing and updating their protocols, assessing the quality of included primary research, using multiple reviewers to screen or code items, and evaluating the potential risk of publication bias (Torgerson, Hall, & Light, 2012; Campbell Collaboration, 2019).

Through advances in communicative technology and the growth of education research as a field, there is currently such an abundance of research information that synthesis approaches are needed to make sense of the evidence. Systematic reviews can highlight trends in theory or show what kinds of interventions have been used and where, as well as identifying where further research may be helpful. In addition, a systematic review can be a necessary step for doing a meta-analysis, or quantitative synthesis, as was done here. The specific metaanalytic approach is detailed in a later section, but meta-analyses can add value for educational stakeholders by showing overall effects on outcomes of interest, as well as considering potential moderating factors. In response to concerns about an inordinate focus on quantitative research in education, it should be noted that systematic reviews can also be done to locate studies for a "narrative synthesis" (Torgerson, Hall, & Light, 2012, p. 227), to be carried out separately or in conjunction with a meta-analysis. A qualitative or narrative synthesis may be more appropriate than a meta-analysis when a rich or nuanced answer to a research question is sought, or when it is inappropriate to statistically combine primary studies. In the current research, while the goal was to estimate a general effect through meta-analysis, a qualitative analysis was done first to avoid potentially privileging studies whose interventions were found to be more effective. The goal of the narrative synthesis was to characterise the included studies and interventions and identify potential moderators of the effect. The importance of this stage for interpreting the quantitative results is enunciated by the editors of *Review of Educational Research* in their advice for undertaking "integrative reviews": "Meta-analyses are of particular interest when they are accompanied by an interpretive framework that takes the article beyond the reporting of effect sizes and the bibliographic outcome of a computer search" (American Educational Research Association, 2022). Thus, the intent of research *synthesis* is not only to combine results from multiple studies, but to integrate qualitative and quantitative perspectives to produce a richer picture of the research area

3.3 Potential concerns for executing a research synthesis

Despite their benefits, synthesis methods are still being refined and have some potential drawbacks for education researchers. Due to their structured nature, systematic reviews and meta-analysis may be resource intensive and require a team rather than a single researcher to execute. Although this was not possible for the current review, I did consult colleagues to double-check the review methods at several stages. Planning the review could also be a challenge for education researchers. Several published guidelines arose primarily for reviews of healthcare studies, and few were aimed specifically at reviews of educational research until recently (see Table 6). Perhaps due to lack of clarity or feasibility concerns, not all reviews adhere to published standards (Atkinson et al., 2014). For example, Kogut et al. (2019) examined 40 systematic reviews of mathematics education and found that few met the Campbell Collaboration standards. Large bodies, like the What Works Clearinghouse (WWC), and Best Evidence Encyclopedia (BEE), and EPPI-Centre do not all have the same inclusion standards or include outcomes in the same ways (Slavin & Madden, 2011) Even journals that publish such reviews do not always require that these guidelines are followed (Tao et al., 2011). Having clear inclusion criteria also has its benefits and disadvantages. Research syntheses have sometimes appeared to exclude most of the available studies because of over-rigid inclusion criteria (Higgins & Hall, 2004). Including primary studies of varying quality could lead to unclear review results, but public confidence in research synthesis could also be undermined by protocols that simply reject low quality studies, without even considering them in a scoping review (Slavin, 2008). In the current review, studies of multi-designs were included, within certain limits, but the requirement of being able to compute an effect size led to numerous exclusions. Thus, the aim of being "exhaustive" on the research topic is potentially impractical. Other challenges regarding the meta-analysis are discussed below.

Regarding the narrative synthesis and identification of moderators, there are fewer "official" guidelines, and it can be a challenge to know what data to extract from the primary studies. As shown below, determining the qualitative categories was a recursive process for the current review, and this is an area where researcher bias could have crept in. Some reviews have chosen more theoretically determined categories, such as the SOLO categories for describing interventions (Hattie, Biggs, and Purdie, 1996) or different models of MC/SRL functioning. With these approaches, the synthesis results could more easily feed-back on theory or future research. On the other hand, theoretically categorised interventions could be more challenging for teachers to operationalise. For example, See, Gorard, & Siddiqi (2016, p. 65) found that teachers, in their efforts to implement effective feedback as defined in Hattie and Timperley (2007), perceived they were already applying the relevant theories to their practice, while the evaluators saw many areas for improvement. Since applicability for teachers was a goal of the current research, I emphasised the common theoretical foundation of the MC/SRL interventions and sought to differentiate them based on more practical elements. Yet this and the other goals of the narrative synthesis were hampered by the reporting of individual studies, especially the paucity of intervention details. The limitations of using the synthesis approach for the current research are discussed further in the final chapter. These issues notwithstanding, the methods chosen were the best ones available to address the research questions, and the results should lead to more focused and directly applicable studies in the future.

Organisation, location, and website	**Subject focus**	**Systematic review guidelines**	**Conducts and reports systematic reviews?**	**Other resources offered**
EPPI-Centre (UK) https://eppi.ioe.ac.uk	Education	General guidance	Yes	Software tools, training, consultation
What Works Clearinghouse (US) https://ies.ed.gov/ncee/WWC	Education	What Works Clearinghouse Procedures Handbook (Conducting and reporting)	Yes	Registers protocols, training
Best Evidence Encyclopedia (US) https://bestevidence.org/	Education	General guidelines, "Best evidence synthesis"	Yes	None listed
Campbell Collaboration (Norway, International) https://www.campbellcollaboration.org/	Social policies or programmes	Campbell systematic reviews: policies and guidelines (Conduct and reporting)	Yes	Pre-registers titles, publishes protocols, training, statistical tools
Cochrane (UK, International) https://www.cochrane.org/	Healthcare	Cochrane Handbook for Systematic Reviews of Interventions (Conduct) MECIR (Conduct and reporting)	Yes	Software tools, training
JBI (Joanna Briggs Institute), (Australia, International) https://jbi.global/	Healthcare	JBI Manual for Evidence Synthesis	Yes	Software tools, training, registers protocols
PRISMA https://www.prisma-statement.org/	Healthcare and other interventions	PRISMA Statement (Mainly reporting)	No	Guidelines and templates for protocols, systematic review reports, flowcharts

Table 6. Sources of guidance for systematic reviews in education.

3.4 Systematic review methods

This section describes the protocol for the systematic review, the systematic search and screening process. In order to demonstrate how the research questions for the review were addressed through clear, systematic, and replicable procedures, these are reported in detail following recent recommendations (e.g., Atkinson et al., 2014).

3.4.1 The review protocol

The review protocol drew most heavily on the PRISMA standards for systematic reviews (Moher, et al., 2009)[13]. While PRISMA is mainly a reporting guideline, it can also be used to structure the review in the first place. I adopted the PRISMA-P (Shamseer et al., 2015) format for the review protocol. Because the process of systematic reviewing was still being learned throughout this research, it was not possible to have a complete, publishable protocol before implementing the review methods, but it would be beneficial to register a protocol for a future, more focused review. The sections of my protocol that were drafted before implementation included PRISMA-P sections 1-8, covering the title, information about the reviewer, funding, rationale and relationship to previous reviews, research questions and eligibility for included studies, and inclusion/exclusion criteria (see Appendix 2 and 3). Search outlets and strings were also drafted, but these went through several versions before being finalised as described below. Changes that needed to be made to the protocol throughout the review process were noted with the date of change, and discussed with the review supervisor, but only the final version is included in the appendices. Noting where methods diverged from the plan is recommended in systematic review guidelines (e.g., Torgerson, Hall, & Light, 2012, p. 220; Campbell Collaboration, 2019, pp. 38-39).

First the type of research to be gathered had to be determined. Studies that could best answer the research questions would have experimental or quasi-experimental designs with planned outcomes in mathematics skills or achievement. To make more convincing causal claims, there needed to be a comparison group (Gorard, 2013, pp. 94-96), and it was intended that allocation to groups should be unbiased, that is, random or probability-based (Campbell Collaboration, 2019, pp. 9-12). This criterion was broadened as it was found the method of allocation to groups was not reported clearly in some cases, and there was a range of grouping methods being described as “random” and some explicitly non-random. With regard to outcome

[13] There is now an updated version of the PRISMA statement, published in 2020, but the design for this review was determined while the 2015 version was current.

assessment, it was expected that pre- and post-intervention assessments would be used, but In practice, not all studies used or reported pre-test scores, and there are arguments for using only posttest scores, especially if there was random assignment to groups to control for potential imbalances.[14] So again the choice was made to include all studies with a non-MC/SRL comparison group and at least posttest data sufficient to calculate a between-groups effect size.[15] If effect sizes were reported in the study without the data needed to re-calculate them, this could be problematic because of the weighting and other aspects of the intended quantitative analysis. With these design requirements set at the beginning of the review, non-empirical or observational studies looking at self-regulated learning or metacognition in mathematics would be naturally excluded. Other designs, such as case studies, could certainly aid understanding of MC/SRL interventions, but they would not be able to support strong, generalisable claims about the intervention effects, and so were not included. These choices were necessary to conduct a coherent review and do not negate the quality or potential value for practice of any excluded studies.

To design a clear and replicable search, the specific words and phrases must be considered, but this fact means the results are constrained by the terminology used in reports. In education studies, there is yet to develop a controlled vocabulary to optimise literature searches, similar to the Medical Subject Headings (MeSH) in healthcare research (National Library of Medicine, 2021). For the current review, the type of included studies was chosen early on, but it was challenging to specify design elements in the search terms due to the myriad ways in which design can be described in education studies. With their greater internal validity (Campbell Collaboration, 2019, p. 10), randomised controlled trials (RCTs) have been praised by some education researchers, while being sharply critiqued by others (e.g., Morrison, 2021; Norman, 2003). Perhaps due to uncertainty around the appropriateness of RCTs in education, the RCT label may not always be used in reports even when one has been done, and this trend was observed in the final review set. On the other hand, broader labels like “quasi-experimental” may be applied to studies without active manipulation, comparison groups, or numeric outcomes, and these omissions would preclude inclusion in the current review. Although some terms related to design were included in the search, it remained the case that design factors led

[14] For example, posttest-only designs would be less susceptible to a test effect.

[15] I intended to use the pre-test data, where available to calculate a more accurate, study-level effect size. Results for both a posttest only, and a pre-post effect are given, with an explanation for why the former is considered the official result.

to the largest number of studies screened out, indicating the potential benefits of higher specification in the searches.

PICOS element	Inclusion criteria	Search terms
Participants	School aged (3-18) students, individually or in groups, in regular mathematics classes.	X
Intervention(s)	Based explicitly on metacognition or self-regulation, delivered in school settings, over multiple sessions.	meta-cogn* metacogn* self-reflect* self-regulat*
Comparator(s)	Normal mathematics teaching or an alternative intervention.	math*
Outcome(s)	Mathematics performance on a validated assessment used at higher than classroom level.	math*
Study Design	Experimental (including RCTs) or quasi-experimental, where the intervention is preplanned and documented, with set, quantitative outcomes sufficient to calculate an effect size.	treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate *)

Table 7. PICOS elements, inclusion criteria, and search terms for systematic review. Green text indicates elements directly reflected in search terms, while red text indicates criteria that could not be directly included in searches.

3.4.2 The systematic search

As shown in Table 7, the search strings were built around the PICOS framework (Campbell Collaboration, 2019, pp. 7-8; Torgerson, Hall, & Light, 2012, p. 220; Shamseer et al., 2015, pp. 7-8)[16]: *Participants, Interventions, Comparators, Outcomes,* and *Study design*. However, not all the PICOS elements were productive to include in the searches. For example, the "participants" category was meant to include only children ages 3-18 in regular school or pre-school settings where mathematics was normally taught, however this could not be limited in the search terms due to the various ways school grades or years are labelled internationally. Similarly, the "comparators" category should have specified business-as-usual or alternate intervention comparison groups, and "outcomes" should have specified only validated assessments of mathematics skills or achievement, not attitudes or other affective outcomes in mathematics. These elements, however, could not be precisely specified in the search terms,

[16] The PICOS/PICO acronym has slightly different interpretations in each source. My interpretation is shown.

without restricting the returns too greatly, so they had to be screened for manually. For the "intervention" category as used in the searches, this was broadly conceptualised to be any treatment based explicitly on metacognition or self-regulated learning, with minimum length requirements of at least 2 hours. The intervention would also need to be documented and replicable to be meaningful for teachers. However, throughout the screening process, it was found that several interventions intended to influence metacognition or self-regulated learning development involved changing the basic structure of the schooling, such as block-scheduling, full- versus half-day kindergarten, or having a teaching aide in the classroom. Because these programmes could not be taken up by mathematics teachers in their normal classes, the conceptualisation of the intervention was refined to be some kind of "training" for students in self-regulated learning or metacognition. This training could still be explicit, such as monitoring strategies during problem-solving, or implicit, such as journaling to boost self-reflection. By narrowing the intervention description, I intended that the group of included studies would have more comparable effects and these would be more applicable to mathematics teachers. Again, this focusing was done following the searches within the screening process.

Several versions of the literature searches were tried out in different search outlets before settling on the final search strategy. Some searches returned only a few items, while others returned an impractically high number. In any review, there is a need to balance sensitivity and specificity, to tailor the searches so that a high proportion of well-fitting items are returned without returning so many ill-fitting ones that the screening process becomes unworkable (Torgerson, Hall, and Light, 2012, p. 224). In refining the searches, it is possible that researchers' personal biases could influence the review outcomes, but the choice to be more specific and have a smaller set of studies to screen and analyse is a necessary limitation of the current review, especially as it was done by a single reviewer and not a multi-member team. The aim was to ensure that the results are representative of the field, if not exhaustive, and that the conclusions are sound (Atkinson et al., 2014, p. 94).

Search outlets feature different capabilities for searching, and some reviewers have chosen to craft different search strategies for each outlet to take advantage of these. For the current review, only very simple operators (i.e., AND to join search terms) were used to keep the searches as consistent as possible between search outlets, some of which do not support complex search term combinations. The basic search string is below and was used to search the text of abstracts within the chosen databases. Keywords, alternate forms, and index terms were not intended to be used, since it was believed that could lead to a high number of irrelevant items, especially given the potential terms that could be related to metacognition or self-regulated learning. These theories needed to be explicitly mentioned in the report text for a study to be included, so it was not intended for the searches to retrieve items that mentioned executive function, for example, without explicitly mentioning one of the two core terms. An exception was made in the case of self-reflection, which is a term also featured in the search. It was expected that studies could feature self-reflection techniques as part of the intervention tested, and mention this in the abstract, while laying out a rationale based on metacognitive or self-regulated learning theories in the full-text document. This choice may have led to more items being returned by the searches which needed to be excluded in the screening process, as there were some studies that were based on self-reflection theories and that made little or no reference to self-regulation or metacognition. The abstract rather than full text was chosen for the search to ensure that the search terms featured prominently in the article. Still, some reports included the two core theories in the abstract, while drawing only tenuous links to them in the full-text reports.

(meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND (math*) AND (treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)

Figure 4. Basic search string for the systematic review.

In choosing outlets to search, several factors were considered. The outlet had to be publicly available or accessible through institutional subscriptions. It needed to allow operators to combine alternate terms within a single search and wildcards to locate different versions of the search terms. The ability to download the citations as a batch was also required, not only for practicality, but also because downloading citations individually could increase the risk of error in the search records. Several institutional librarians were consulted before the systematic review searches were implemented. They were able to point out where databases overlapped, for example, but were not able to give detailed guidance on the design of the searches or the most appropriate databases to search. However, my university library does have a web page regarding systematic reviews (Bisset, 2020), which mainly echoes other sources on systematic review best practices. Each database or search outlet also offers differing levels of advice on how to optimise a search. It is clear that not all databases are set up to handle the demands of systematic searching and screening, as some offer limited options for combining search terms or downloading references in batches rather than one at a time. In some cases, a search outlet was excluded after trying out searches and discovering this lack of functionality. It should be noted that it is possible for databases or search outlets to change their publication coverage over time, and these changes could be backdated. This means that even when searching a set range of dates in the past, the search returns could vary slightly if the search is re-run after some time. No systematic search is ever likely to be fully replicable. Appendix 4 displays the search outlets included in the final search and the number of returns, and duplicates that were removed later. Appendix 5 shows sources that were considered for the review but ultimately rejected, with reasons.

3.4.3 Screening for inclusion

Once the final searches had been run, the references from each search outlet were added to a review file in the EPPI-Reviewer software developed by the Institute of Education at the University of London (Thomas et al., 2020). It should be noted that this software has gone through several versions, but EPPI-Reviewer 4 and EPPI-Reviewer Web were the versions used for this review. EPPI-Reviewer offers many functions to assist in the systematic review and meta-analysis process, but manual adjustments can still be needed, for example in the deduplication process. In EPPI-Reviewer, the automatic deduplication feature can have its sensitivity adjusted to mark items that have a greater or lesser degree of similarity to each other as duplicates. When the same item was returned by multiple searches, the authors' names, titles, abstracts, and publication type were often listed differently, leading to a lower similarity score. If the program is uncertain about the duplicate status, it requires manual approval of a set of items as duplicates. Many duplicates were missed by the automatic screening and needed to be manually added to the duplicate list throughout the screening process.[17] The numbers here reflect the final counts. Out of the total returns from the searches, more than half (1,877) were found to be duplicates, while 1,761 unique items remained for screening by title and abstract.

> However careful the search was, there will be reports that do not contain any research evidence, ones that are unclear about what was done and found, and some that are actually irrelevant to your area of interest. These can be eliminated from further consideration by skimming the abstract. (Gorard, 2013, p. 28)

There was a multistage process of screening the studies for inclusion in the review, as recommended by Torgerson, Hall, & Light (2012, p. 226), which began by considering the titles and abstracts. This screening may be represented as straightforward, as illustrated by the quotation above, but the reality is more nuanced. Having preset inclusion criteria decreases the need for reviewers' judgement at this stage, but it does not eliminate it because abstracts can omit key information about a study's design, sample, assessments, and even academic subject area. When there was insufficient detail in the abstract to exclude a study, it was retained to the

[17] Manual de-duplication was made more difficult by the fact that research items from the same author(s) can have very similar titles or abstracts. DOIs could have been used as unique identifiers but were not listed for all items, and it was not possible to sort items by DOI to check for duplicates. It is possible that the set of items excluded from the review includes missed duplicates or that some items were wrongly marked as duplicates.

next stage of screening on the full-text report. Table 8 shows the screening codes applied to each title and abstract and how many studies received each code in the first screening stage. Note that some items were excluded for multiple reasons, so the number of codes does not sum to the number of items. The exclusion codes were determined based on examples from methods literature (e.g., Torgerson, Hall, and Light, 2012, p. 223), guidance from my supervisor, and the unique needs of my review approach.

Table 8. Frequencies of applied codes after screening on titles and abstracts, with finalised descriptions of each code.

Code	**Items Coded**	**Coding Description**
Include Confidently	135	Include based on title and abstract. Need to retrieve full report for full text screening.
Include with Reservations	279	Unsure about this one due to lack of information in title and abstract. Carry forward to next stage of screening and confirm the study details noted here with the full text.
Exclude on Date	31	The study was published before 2005 or after 2019.
Exclude on Language	2	The study is not available in English.
Exclude on Publication Type	96	Exclude: introductions; monographs, books, and book chapters; encyclopaedia, dictionary, or other reference works (or entries), journals, magazines, or websites not subject to peer review; works not listed in common academic databases; unpublished studies or data sets.
Exclude on Topic	230	Exclude any study not regarding academic teaching and learning, even if metacognition (or similar) and/or mathematics are mentioned. Also exclude studies about research design per se.
Academic Subject	74	Exclude if mathematics was not taught or evaluated as an intervention outcome. Do NOT exclude if other subjects are taught or evaluated as outcomes alongside mathematics. Do NOT exclude if general academic achievement or skills is mentioned, without giving a specific subject.
Exclude on Population Age / Level	284	The subjects of the intervention are NOT students aged 3-18 in primary or secondary settings. Exclude studies where teachers or teacher-trainees, college or university students, parents, or infants under three years of age are primary focus of the intervention/training or of the outcomes assessment.
Exclude on Sample Size	32	Exclude any study with less than 10 total participants from the same population age/level.
Exclude on Research Setting	39	The study was done in a laboratory or other setting in which teaching mathematics is not a normal activity. Also exclude settings in which mathematics teaching is done intensively as a primary or only activity, such as summer schools or tutoring clubs. Also exclude studies in which the intervention or assessment was partly or completely done at home.

Code	**Items Coded**	**Coding Description**
Exclude on Design	739	The study is NOT an experiment or quasi-experiment, in which at least one aspect of teaching is manipulated by the researchers and some type of comparison group is included. Exclude when participants select their group or level of intervention. Exclude protocols/plans that do not report on work actually conducted. Exclude review studies. Exclude purely measurement/observation or assessment validation studies.
Exclude on Intervention / Treatment	39	The intervention is not training for students based on a stated metacognition/self-regulated learning approach (even if these are assessed). The intervention period is less than at least 2 hours or two sessions long. Also exclude interventions that are embedded into the outcome measures, with no other intervention or assessment.
Exclude on Outcomes / Evaluation	11	Mathematics learning or achievement was not a designed outcome. Exclude if outcomes were not measured in a structured and replicable way. Exclude if only observational, qualitative, or non-numeric outcomes are used, or there is insufficient data to compute an effect size.

Table 8 (cont.). Frequencies of applied codes after screening on titles and abstracts, with finalised descriptions of each code. Each item was given one or more codes. Code frequencies do not sum to the number of items screened.

Creating the exclusion codes necessitated, in some cases, clarification of what was required for inclusion. For example, the review aimed to identify MC/SRL interventions that would align with the Campbell Collaboration's category of "Generic types of programs or practices [. . .] not limited to a brand name version" (Campbell Collaboration, 2019, p. 8). Originally, this was specified as any programme explicitly linked to MC/SRL theories and having mathematics outcomes, but this needed to be slightly focused during the screening process. Some studies mentioned self-regulation/self-regulated learning or metacognition as part of the rationale for the study or as measured outcomes, but the interventions relied on substantially changing the structure of the school day, such as by implementing whole-versus half-day kindergarten or using a whole, branded curriculum. These programmes were too dissimilar to interventions that teachers could implement in mathematics class. Therefore, the code for "exclude on intervention/treatment" was applied when the programme could not be conceptualised as direct or indirect "training" for students in metacognitive or self-regulated learning skills. The code, "include with reservations," also needs comment: many reviews use only one "include" code in the screening stage, while the rest account for exclusion reasons. Because of the ambiguity found in some abstracts, "include with reservations," was used to tag items that needed specific aspects to be checked in the full-text version, while "include confidently" was used for items where every aspect of the study reported in the abstract seemed to fit the criteria well.

While the screening on title and abstract was finalised by me alone, second raters assisted at two points. First, my thesis supervisor joined in screening 10 items early in the screening process. This was a collaboration undertaken to familiarise myself with the process of screening on titles and abstracts. A formal interrater reliability was not calculated for the discussions of items, but the supervisor's expertise in systematic reviews aided the development of the screening protocol. I also took advantage of this expertise through consultations over email or video call regarding specific aspects of the studies being screened. Next, a fellow doctoral student at my university, also undertaking a systematic review, was enlisted to screen 109 abstracts[18], and his coding was compared to my own. Overall, there was full agreement on 61 items and some difference in the coding on 48 items, mostly regarding the reasons for exclusion. There were only nine disagreements about the actual include/exclude decision. For two of these, I had chosen to exclude them, while my colleague included them, but for the other seven items the reverse was true. Regarding the specific codes applied, there were

[18] There were 110 items in the intended double-screening list, but one item was found to be a duplicate and was therefore not coded.

more differences, especially given that each item could receive multiple codes. In some cases, my colleague used the "wrong" code, such as excluding on "topic" (i.e., not about teaching and learning in schools) when the "academic subject" (i.e., not about mathematics learning) exclusion code would have been better aligned with the coding descriptions. Based on considering the double-screening, the most logical response was to maintain the codes I had originally applied, rather than change some and risk losing comparability with the other items. Assuming the two items I excluded but my colleague included were representative of a larger set of "mis-coded" items, there would be about 33 additional items out of 1814 which should have been retained for screening on the full-text report, but it is unlikely many of these 33 would be used in the final review, most likely around three or four items.[19] To reiterate, I took the final include/exclude decision at each stage of the review.

Moving on to the full-text screening stage, it was necessary to retrieve the reports from the publishers, the authors themselves, or personal contacts. Fifteen reports were not available, of which several were also found to have other reasons for potential exclusion, such as being listed as a book chapter or being published in a language other than English. Ideally, these other aspects would have been confirmed with the full reports, but this was not possible. Logging into the university library allowed me to access several publications that would otherwise have charged a fee. For a handful of studies in this review (four reports), access was not possible without paying additional fees, and these studies were therefore excluded without regard to the specific amount to be charged. PDF files of the full-text reports were uploaded to the EPPI-Reviewer website where they could be read and annotated. Because so many potentially fitting items were identified in the earlier screening, practicality considerations partly determined the full-text screening. Rather than reading every report straight through, I read purposefully, examining each section of the report that might contain information necessitating exclusion of the study, such as the descriptions of the population/sample, research design, intervention, setting, and numeric results. Introductions and discussions/conclusions were not likely to yield exclusion-necessitating details and were usually only glanced over. While the procedure for screening on titles and abstracts retained all studies when there was uncertainty, I excluded studies immediately upon finding any aspect from the full-text screening that did not clearly fit the criteria. However, because I was still learning about the many ways that effect sizes can be generated for meta-analysis, I retained any reports that had appropriate designs

[19] Because these would be items about which I had reservations from the title and abstract screening, I applied the final acceptance rate of items in this category after the full-text screening, which was 11%. For items coded as "include confidently" in the title and abstract stage, the final acceptance rate was about 49%, which would indicate 16 or 17 as the maximum number of wrongly excluded items.

but reported numeric results in an unexpected format, until the numeric data could be evaluated further. Some items considered “borderline” were discussed with my supervisor before I made the final decision to include or exclude them. The code frequencies applied to the full-text reports are in the tables below.

Table 9. Frequencies of applied codes after screening on full-text reports, with finalised descriptions of each code.

Code	**Items Coded**	**Coding Description**
INCLUDE on Full Study	62	Include based on full-text report. Item ready for in-depth review.
Exclude on Full-text Not Available	15	Full text is not publicly available for the study report. Exclude conference and other papers that are shorter than 1-page (250 words). Exclude where the electronic files for full-text papers are not available after searching Google and Durham library holdings, using personal connections, and contacting authors by email or other electronic means. DO NOT exclude for full-text versions from ERIC or sent through email, following an effort to procure the "official" version of the report.
Exclude on Date	4	The study report has an official publication date before 2005 or after 2019. Exclude studies posted online in 2019, but officially published in 2020 or later. Include studies posted online in 2004, but with an official publication date of 2005 or later.
Exclude on Language	32	The study is not available in English. Exclude studies that have an English abstract, but the full text is in any other language.
Exclude on Publication Type	10	Exclude: introductions; monographs, books, and book chapters; encyclopaedias, dictionaries, or other reference works (or entries), journals (e.g., trade journals), magazines, or websites not subject to peer review; works not listed in common academic databases; unpublished studies or data sets. DO NOT exclude post-graduate theses, or conference papers longer than 1-page (250 words), as long as they are the "accepted" versions. DO NOT exclude chapters/articles from "books" that are actually conference proceedings.
Exclude on Topic	0	Exclude any study not regarding academic teaching and learning, even if metacognition (or similar) and/or mathematics are mentioned. Also exclude studies about research design per se.
Exclude on Academic Subject	8	Exclude if mathematics was not taught or evaluated as an intervention outcome. Do NOT exclude if other subjects are taught or evaluated as outcomes alongside mathematics.

Exclude on Population Age/Level	30	The subjects of the intervention are NOT students aged 3-18 in primary or secondary settings. Exclude studies where teachers or teacher-trainees, college or university students, parents, or infants under three years of age are primary focus of the intervention/training or of the outcomes assessment.
Exclude on Sample Size	6	Exclude any study with less than 10 total participants from the same population age/level. Even if there are more than 10 participants overall, exclude if less than 10 participants completed the outcome assessments used to compute effect size.
Exclude on Research Setting	45	The study was done in a laboratory or other setting in which teaching mathematics is not a normal activity. Also exclude settings in which mathematics teaching is done intensively as a primary or only activity, such as summer schools or tutoring clubs. Also exclude studies in which the intervention or assessment was partly or completely done at home, including using homework as a major part of the MC/SRL training.
Exclude on Design	108	The study is NOT an experiment or quasi-experiment, in which at least one aspect of teaching is manipulated by the researchers and some type of comparison group is included. Exclude when participants select their group or level of intervention. Exclude protocols/plans that do not report on work actually conducted. Exclude review studies, essays, opinion pieces, or general reflections on practice. Exclude purely measurement/observation or assessment validation studies. Exclude if there is no non-MC/SRL comparison group.
Exclude on Intervention /Treatment	102	The intervention is not training for students based on a stated metacognition/self-regulated learning approach (even if these are assessed). The intervention period is less than at least 2 hours or two sessions long. Also exclude interventions that are embedded into the outcome measures, with no other intervention or assessment. Also, exclude programmes that represent entire curricula ,not interventions (e.g., Montessori, Tools of the Mind). Exclude if underreporting makes the main intervention elements unclear.
Exclude on Outcomes/ Evaluation	68	Mathematics learning or achievement was not a designed outcome. Exclude if outcomes were not measured in a structured and replicable way. Exclude if only observational, qualitative, or non-numeric outcomes are used, or there is insufficient data to compute an effect size.
Keywords Missing from Abstract	5	Exclude if none of the intended keywords is in the abstract: “self-regulation,” “metacognition,” or “reflection.”

Table 9 (cont.). Frequencies of applied codes after screening on full-text reports, with finalised descriptions of each code. Each item was given one or more codes. Code frequencies do not sum to the number of items screened.

As shown in Table 9, full-text screening utilised the same exclusion codes as the title and abstract screening, but the descriptors of some codes were adjusted slightly to take account of unpredictable study elements in the reports and to ensure the codes were applied consistently. For example, several studies had online publication dates that were within the acceptable range, but their official publication date was outside of it. The wording of the "exclude on date" code was therefore adjusted to clarify that official publication dates only would be considered. The next code, "exclude on language," had a note added because several studies were found to have English abstracts but not full reports in English. The "exclude on publication" description was revised because in some cases, especially with regard to conference papers, it was unclear whether the item was a full-report or long abstract. A minimum of 1 page or 250 words was established to consider a report "full" within the "publication type" category, but it could still be excluded for missing information or other reasons. "Exclude on Topic" was retained from the previous screening stage, yet this code was not applied to the full-text reports because I was able to screen out all studies not about teaching and learning in schools at the earlier stage.

In some other reviews, such as Higgins et al. (2004; 2005), the review team considers a much larger set of studies for a scoping or "mapping stage" review than is finally included in the meta-analysis, if one is performed. The advantage of doing this is to be able to describe the research that has been done qualitatively, even if a meta-analysis is not appropriate or excludes some of the studies. Wang and Sperling's (2020) recent review of metacognitive interventions in mathematics stops short of doing a meta-analysis, but it shows the wide variety of theoretical and practical approaches to this area. The current review seeks to address the research questions in a way that is useful for practitioners, and to carry out the review without greatly adjusting the original criteria, such as the study design. The original intent was to provide an estimate of the overall effect of the target interventions, to aid teachers in choosing approaches to consider. A scoping review which disregarded the numeric outcomes of the interventions might be of interest to stakeholders, but it would not clarify which interventions had more promise to boost learning. In addition, this review already includes more primary studies in the meta-analysis than other comparable reviews, but it would be impractical for me as an independent researcher to do a larger, qualitative review with less stringent inclusion criteria.

3.4.4 Narrative synthesis methods

Once the list of included studies was finalised, I began the data extraction from the full study reports within EPPI-Reviewer Web, which has both “checklist” and “line by line” coding functions, as well as the option to enter descriptive notes for any code selected. I carried out the data-extraction independently in a detailed, three-stage process, with periodic supervisory discussions regarding uncertainties in the study reports. While the intention was to estimate the general effect of MC/SRL studies in a meta-analysis, I decided to undertake a qualitative synthesis prior to extracting the quantitative information needed to generate the effect size. As stated earlier, there was a concern that the qualitative synthesis might overlook important aspects of the studies and interventions if it was known which studies had higher effects, and the qualitative data were extracted and analysed first for this reason. As with the inclusion/exclusion codes, the data extraction categories were determined based on consulting methods literature, discussions with review experts, familiarity with similar MC/SRL trials, and my own research questions. Like a “grounded theory” approach (Bryman, 2008, p. 543) in qualitative studies, some of the categories and codes were developed while reading the reports themselves based on salient potential trends and themes. Some intended categories were shifted or eliminated during the coding due to inconsistent reporting. In this way, the qualitative synthesis was potentially more open to researcher bias than the meta-analysis, and the results should be viewed as exploratory rather than definitive. Optimally, such findings would be confirmed through a more focused review done with a team of researchers to permit double-coding of all the reports.

Overall, the qualitative synthesis aimed to describe several key aspects of the included studies: the educational contexts and participants, the study designs, the MC/SRL interventions, any ethical issues, and the quality of the evidence from the studies. Several of these items align with the TIDieR framework for intervention reporting (Hoffmann et al., 2014), but they were tailored to the MC/SRL focus of this research. The full codes are given in Appendix 6. In addition to describing the interventions and study designs, I also intended to uncover fruitful ways of analysing potential moderators of effect within the meta-analysis to answer the second research question. As will be seen in the results section, the qualitative synthesis and moderator analyses were limited by issues with reporting in the study reports. With regard to the interventions, several methods of categorising MC/SRL training have been used in previous reviews. Some reviews have categorised studies based on whether they used cognitive, metacognitive, or motivational theories or strategies (e.g., Dignath, Büttner, & Langfeldt, 2008;

Wang & Sperling, 2020). However, it is not clear how interventions were sorted into these categories, and many interventions in this field use multiple strategies. Therefore, for the current review, I decided to categorise interventions based on the specific activities or MC/SRL strategies they taught and by the SRL "stage." I did not compare MC/SRL interventions based on different theoretical models, partly because it seems that students and possibly teachers were not thoroughly oriented to such models, and partly because all included interventions were thought to operate through the same feedback cycle of metacognitive knowledge and control.

3.4.5 Methods for posttest-only meta-analysis

Following the preliminary narrative synthesis, the quantitative information used to compute study-level effect sizes was extracted from the included reports into a spreadsheet (see Appendix 7). At the time of the review, the current version of EPPI-Reviewer (ER-Web) did not offer the facility to undertake a meta-analysis, so a separate spreadsheet was used to allow porting the data into another meta-analysis tool. Google Sheets were used to take advantage of the simplicity of the interface, ease of sharing, and automatic back-up. Effect sizes were computed for each study using the effect size calculators from the Campbell Collaboration (Wilson, n.d.). These tools were chosen for two main reasons: First, calculating effect sizes by hand was believed to be more prone to human error and inconsistency. Second, using these tools allowed an effect size to be generated for each study, even where the statistical reporting differed, in a more straightforward way than would have been possible using a comprehensive meta-analysis tool. These calculators generate effect sizes based on formulas from Lipsey and Wilson's (2001) *Practical Meta-analysis*. The Hedges' *g* effect size is given to four decimal places, and the correct sign is applied when the comparison group outperforms the treatment group. Hedges' *g* is considered an "unbiased" effect size estimate because it corrects for bias due to small samples (Lin & Aloe, 2021), yet in many cases Hedges' *g* is very similar to the more commonly used Cohen's *d*, and in some cases, it can produce a slightly larger value. For the current review, differences between the two values were around 0.01 and thus not considered meaningful. These calculators also return the variance for the effect size and a minimum and maximum value for the 95% percent confidence interval. The variance was used to compute the standard error needed for meta-analysis.

In the first round of calculations, effect sizes were generated based on performance on the main mathematics assessment at posttest (i.e., the first assessment point following the end of the MC/SRL training)[20], as all included studies had these outcomes. When multiple mathematics assessments were used, I chose only one as the main outcome for this review based on whichever was the most objective, broad-scope measure. Group sample sizes at posttest, along with unadjusted group means and standard deviations, were the preferred data to extract for the initial effect size estimate, though in some cases these needed to be reconstructed or interpreted from the description in the report. For example, not all studies reported the sample sizes at each assessment point, so outcome samples were assumed to be the same as initial samples unless otherwise indicated. Whenever the reporting was ambiguous or contradictory, the smaller of the two numbers was assumed to avoid overestimating the study-level effect size or its weight in the summary estimate of effect. Similarly, if only the total sample was given and not the sample for each group, then the smallest, equal-sized groups adding up to the total with interval numbers of pupils were assumed unless the description indicated otherwise[21]. See Appendix 8 for a worked example of how this assumption could have impacted the study-level effect. Some studies only reported outcomes by subsets of the intervention or control groups. In these cases, the "subgroups" Campbell Collaboration calculator was used. This calculator was also used whenever there were multiple MC/SLR interventions to be combined into a single, study-level effect size estimate, and when there were groups of multiple ages or grade-levels that were taught separately using the same intervention and assessed with the same test (e.g., Vula et al., 2017).[22]

Following the calculation of study-level effects, data was entered into Meta-Essentials (Suurmond, van Rhee, & Hak, 2017), which operates as a special workbook within Excel. This programme generated the combined effect and forest plot of included studies, as well as the heterogeneity statistics and analyses for potential bias (e.g., publication bias). All outcomes from Meta-Essentials are reported without any adjustment or correction in the next chapter, beginning with the posttest-only meta-analysis.

[20] The first assessment of the chosen mathematics assessment following the assessment period was considered as the "posttest," regardless of the original authors' terminology. Any test done during the intervention period was ignored.

[21] This was done to avoid "cutting a student in half" since every student could only belong to one group. When the reconstructed groups were added together, this could result in a very slightly smaller total sample than that reported, but this is noted in the table where it occurred.

[22] When different age/grade-level groups were given different assessments, then an effect was computed at each level and these were combined in a fixed-effect meta-analysis to generate a single, study-level effect for use in the larger random-effects meta-analysis.

3.4.6 Pre/post and "best guess" meta-analyses methods

While the first objective was to generate an effect from comparing only the posttest scores of the group(s) based on MC/SRL to the non-MC/SRL comparison group, this could be unreliable if there were notable differences between the groups at baseline, and this could potentially lead to increased heterogeneity in the meta-analysis. Such differences are more likely when there is a lack of individual-level random allocation to groups. While many studies in this review reported using randomisation, this was usually done at teacher, classroom, or school level, and it was sometimes combined with matching or other efforts to ensure "balanced" groups. It was therefore considered optimal to include the pretest data, if reported, to compute a more robust combined effect size.[23] Pretest values were not used if the assessment tool was substantially different to the posttest, but it was not considered problematic if different forms of the same test were used. Not all studies reported pretest scores or gains scores, even when a pretest was done. In some cases, authors only reported whether pretest differences between groups were "significant," without reporting the group means and standard deviations.[24] In order to calculate a pre/post effect size, the mean gain (i.e., the difference between the pretest and posttest means, which could be negative), was entered into the relevant Campbell Collaboration calculator, along with the pre- and posttest standard deviations, and the alignment between the pre- and posttest scores, either *r* or the results of a *t*-test.

Because multiple assessments done with the same participants cannot be assumed to be independent, it is necessary to have a measure of the alignment between the observations at different timepoints. However, most studies reporting pretest or gains scores did not report this correlation. In such cases, Borenstein et al. (2009, pp. 232-233) suggest it is possible to estimate the correlation, but that it is necessary to consider its impact on the meta-analysis. Although setting the correlation to different levels does not impact the effect size itself, it does impact the variance and standard error, which in turn affects the interpretation of the heterogeneity, for example. If we assume a high correlation between the pre- and posttest scores of the same individuals, then the variance and standard error will increase. One interpretation is that there is less certainty around the effect size because the outcome is more a result of within-participant factors and less a result of the intervention. Setting the correlation to 1.0, for example, would imply that the group conditions had no effect on outcomes, while setting

[23] It is worth noting that if pre-test SDs were smaller than those for the posttest, including pretests would inflate the ES calculation, even if there was baseline equivalence between the groups.This is because the ES unit, the pooled SD, is smaller.

[24] Pre-test differences were also generally reported as "significant" or not, without reporting the actual degree of the difference between treatment and comparison groups.

it to 0.0 would be akin to treating the pre- and posttests as completely independent measurements. Neither is considered appropriate here. Cole et al. (2011) state that pretests are ". . . typically the single best covariate for explaining posttest variation . . ." (p. 2), and they point out that psychometric properties of the test itself, such as measurement error and discrimination across the ability range, can limit the correlation. Using large-scale assessment data from several US states, Cole et al. (2011) demonstrate that pre/post correlations varied depending on the testing location and properties of the students sampled, such as ability. Higher- and lower-ability subgroups show decreased correlations relative to the whole sample (p. 7). On the other hand, they found that correlations did not vary as much by the subject area (e.g., mathematics vs. English) or by participants' ages. In addition, when scores are more closely grouped together the correlations are lower because there is more tendency for individuals to switch ranks between pre- and posttest. Cole et al. (2011, p. 35) found an average population correlation of 0.81, SD=0.09, while the lowest performers had a correlation of 0.60, SD=0.13. However, Cole et al. (2011, p. 35) found the minimum correlations ranged from 0.30 to 0.56.

For the current review, many of the included studies used researcher-developed tests that had not been validated on a large sample or were more criterion-based and showed very low pretest scores[25]. In addition, because the scale of many included studies was small, and students were drawn from only a few classes and schools, there is reason to believe the pre/post correlations might be reduced. It was decided to test the impact of choosing different correlation estimates for the pre/post meta-analysis. The correlations tested were 0.5, 0.6, and 0.8, based on the range of correlations for the population and low-performers subgroup in Cole et al. (2011). The studies included were all those studies that reported a pre- and posttest mean and standard deviation and had a single treatment and comparison group, as the effect size calculators used do not permit multiple subgroups within a pre/post analysis. This sensitivity analysis also excluded several studies where only posttest scores adjusted for baseline were reported, as there would be no way to include the pre/post correlation. In all, 30 studies were included, which is a substantial reduction from the original 60. Effect sizes and standard errors were generated for each study based on the 0.5, 0.6, and 0.8 pre/post correlation, and these were entered into separate workbooks in Meta-Essentials (Suurmond, van Rhee, & Hak, 2017). Table 10 displays the outcomes of this sensitivity analysis. It is clear that differences in the combined effects, standard errors, and confidence intervals are negligible or non-existent with

[25] For example, Baliram & Ellis, 2018, p. 99 report a suspected floor effect in the pretest and ceiling effect in the posttest. They also report a pre/post correlation of *r* = .47, which is lower than the average reported by Cole et al. (2011).

all three correlations. There is little difference in the markers of heterogeneity between the 0.5 and 0.6 correlation analyses, with Q and I² being marginally higher with the 0.6 correlation, while T² and T are equivalent. With the 0.8 correlation, Q is noticeably higher, while I² is somewhat higher, but T² and T are slightly smaller. Overall, the greatest impact of increasing the pre/post correlation is that it appears the variation between the effects of the study is greater because the estimates of the within study variance are narrower. It is worth noting that with a 0.5 correlation, the variance and standard error of the effect sizes is almost equivalent to that of the posttest-only effect sizes, while those with the 0.8 correlation are much smaller. Based on these checks, it seemed that using a 0.6 correlation would be preferred, since this indicates that more information is added to the model through using the pre-tests, while not reducing the variance and standard error considerably and thus overestimating the between study variation in effects. In only a few instances, the authors reported the pre/post correlation, or this could be computed from raw data (Edwards, 2008, p. 87), and the actual correlation was then used instead of the assumed 0.6.

Correlation of pretest and posttest scores	**0.05**	**0.06**	**0.08**
Combined effect size	0.52	0.52	0.52
Standard error	0.11	0.11	0.11
95% confidence interval	0.30 to 0.73	0.30 to 0.73	0.30 to 0.74
Q	247.74	297.76	505.20
p_q	0.000	0.000	0.000
I^2	87.49%	89.59%	94.06%
T^2	0.19	0.19	0.18
T	0.44	0.44	0.43

Table 10. Combined effects and heterogeneity with 0.5, 0.06, and 0.08 correlations. RE meta-analysis of pre/post outcomes.

Having made this choice, I thought it appropriate to add back the studies that had been removed for not reporting the pre-test scores or pre/post correlations, but that did report posttest scores adjusted for baseline instead of or in addition to unadjusted scores. When only adjusted outcomes are reported, it is presumed that the “posttest only” effect size calculation is comparable to a pre/post calculation, so these values from the original meta-analysis were added to the workbook. In some cases, this meant that studies with multiple groups were no longer excluded. As with all the analyses, the goal was to utilise all the information in a fair way to produce the best estimate of the combined effect. I also calculated a pre/post effect size for studies that used multiple treatment or comparison groups by combining those into a single effect size to use in the meta-analysis. In primary studies, it is common practice to report a separate "effect" for each intervention compared to the control group or other intervention groups. However, it is inappropriate to combine such effects in a meta-analysis, since this would artificially increase the sample size from each study, giving it undue weight in the final combined effect (Borenstein et al., 239). The goal is that outcomes from each participant should only be used once in the study-level effect and meta-analysis. In some but not all of the primary studies, authors reported outcomes by intervention group and also reported one for "any treatment" (e.g., McClelland et al., 2019; Dresel & Haugwitz, 2008). In such cases, these combined outcomes were used for the pre/post analysis of effect. When outcomes were only reported separately, this was problematic, given that the "gain score" effect size calculator from the Campbell Collaboration does not allow for subgroups within the treatment and control.

Approach	**Explanation**	**Effect Size (*d*)**	**Standard Error**
1	Separate pre/post effects (0.6 pre/post correlation assumed) are computed with each treatment relative to the control, reducing the control sample proportionately. These effects are then entered into a fixed-effect meta-analysis to produce a single ES for the whole study.	0.50	0.16
2	Pretest and posttest M/SD are averaged for all treatments, weighted by sample size. These are then used in a pre/post ES calculation (0.6 pre/post correlation assumed) compared to the control.	0.54	0.16
3	Mean gains (posttest - pretest) and averaged SD (pretest +posttest/2) are entered as subgroups in a single outcome effect size calculation. Pre/post correlation is ignored.	0.54	0.18
Authors' approach	Helmert interaction contrasts, comparing intragroup changes in both treatments, combined, with those of the control.	0.52	0.21

Table 11. Approaches to combining groups in a pre/post effect size calculation, using data from Dresel and Haugwitz (2008).

Three possible procedures were tested for combining subgroups in the pre/post analysis, shown in Table 11 with data from Dresel and Haugwitz (2008) to illustrate potential differences. This example was chosen because the combined results could be compared with those given by the authors (p. 12), the Ψ (*psi*) and SE(Ψ) values reported as an effect size "comparable to Cohen's *d*" (p. 13). Very similar effect sizes and standard errors resulted from these three approaches, and all resemble the authors' own calculation of effect. All approaches make assumptions that cannot be tested without access to the original, individual-level data. Given this, the choice was made to utilise approach 1, in which each group, treatment or control, was separately compared to the other, with sample sizes reduced proportionately. For example, if a study used three MC/SRL treatment groups and a single control group, the size of the control group would be reduced by two thirds for each comparison. If two non-MC/SRL groups were used and a single MC/SRL treatment group, then the size of the latter would be cut in half for each comparison.

After generating an effect size for each pre/post comparison, all the effects were entered into a separate fixed-effect[26] meta-analysis using Meta-Essentials, to produce a combined, study-level effect. There were several special cases, however. In one case, Edwards (2008), individual-level outcomes were included in the report (p. 87), and I calculated means, standard deviations, and pre/post correlations to compute a pre/post effect size. In another case, for Pennequin et al. (2010), F-values were used for the posttest-only ES calculation, but estimated means and standard deviations reported graphically (p. 211) were used to compute a pre/post effect size, with comparisons being made between experimental and control students in the "low" and "normal achievers" subgroups. Kramarski, Weisse, & Koloshi-Minsker (2010) also reported effects by achievement subgroups, so these comparisons were made first, then effects were combined into a single, study-level effect. The same procedure was used for Vula et al. (2017), which used both third and fifth grade classes in the intervention and comparison groups, and for Mevarech et al. (2010), which used third and sixth grade classes. Finally, Wijaya et al. (2018) presented outcomes by test order. Half of participants in each group received form A for the pretest and form B for the posttest, while for the other half this order was reversed. Same test order groups were first compared, then these effects were combined to yield a single effect for the study. Study-level effects were then entered into the combined pre/post meta-analysis, with measures of heterogeneity and potential bias being examined as in the posttest-only meta-analysis. Fifty out of sixty included studies reported either pre/post scores or adjusted posttest

[26] This was chosen rather than a random-effects model because conceptually each study should reflect a single effect of MC/SRL training, even if this took different forms.

scores, and these appear in the pre/post meta-analysis reported in the next chapter. However, to utilise information from all studies, I next executed a meta-analysis (i.e., a “best guess” meta-analysis) that added back the posttest-only effect sizes from the remaining 10 studies. In the results, I explain why neither the pre/post or the “best guess” meta-analysis produced superior results, and I consider the posttest-only combined effect as the official result of the meta-analysis.

3.4.7 Exploratory analyses

To address the second research question, I explored several potential reasons for different effects in the included studies. Coding for these analyses is shown in Appendix 11. In line with Burke et al. (2015, p. 4), these analyses are considered more “hypothesis generating” than “hypothesis testing” due to limitations in the data, but they were based on theory and previous research indicating likely divergences in effects. As discussed earlier, previous reviews have noted differences in the theoretical bases for the MC/SRL interventions (e.g., Dignath, Büttner, & Langfeldt, 2008; Wang & Sperling, 2020) or in their structures. Other differences in the MC/SRL studies, such as in the ages or grade-levels of the participants and the length or “dose” of the interventions, might be correlated with higher or lower effects. In my own review, I wanted to be able to make comparisons with these previous reviews, but I stopped short of drafting hypotheses regarding the potential correlations with effects because I anticipated that reporting and other factors would limit the trustworthiness and meaningfulness of this stage. In addition, the systematic search was not designed around such categories, but it was instead aimed at estimating an overall effect. As shown in the next chapter, the results regarding the apparent heterogeneity of included studies made it even more desirable to explore potential moderators, but the results of the moderator analysis were still mixed. This is described further in the discussion and conclusion chapter.

Chapter 4: Results

In this chapter, I report on the results of the systematic search and screening process, the qualitative aspects of the included studies, and the meta-analysis of the effects on mathematics skills outcomes. First, the results of the qualitative analysis of included studies is presented, and next the results of the meta-analysis. The first section illuminates the types of MC/SRL interventions and study designs used, while the latter shows the average effects on mathematics outcomes.

4.1 The narrative synthesis results

In this section, I present the results of the narrative synthesis exploring the types of MC/SRL programmes evaluated with mathematics outcomes for the years 2005 to 2019. My goals were to outline the included studies, where they were done, and how they were reported, as well as the types of schools, teachers, and pupils involved. Next, I sought to uncover key aspects of the MC/SRL interventions, such as their activities, materials, schedules, connections to mathematics teaching, and other implementation factors, as well as the designs and methods of the included studies. Understanding participants' views of the programmes was also a major goal. All these elements were analysed to form a picture of what MC/SRL training looks like in practice, how clear the evidence is about its effectiveness, and what factors could influence outcomes. As shown below, each of these goals was hampered to some extent by reporting issues, and this has implications for interpreting the results of the meta-analysis of effects that follows in the second half of this chapter. This theme is returned to in the discussion and conclusions chapter.

4.1.1 Included reports

Of the original 1,761 unique items following de-duplication, 62 reports, representing 60[27] separate studies, were retained into the data extraction phase. This process is shown in the PRISMA flow diagram in Figure 5. Two sets of two reports were included that overlap–that is, they discuss the same studies. In one case, there is a dissertation (Schmitt, 2013) and a journal article (Schmitt et al., 2015) that present the same study results, and the two reports differ very little. In the other case, it appears that the same study has been reported in two different ways:

[27] Two sets of two reports each were found to substantially overlap, and this is noted in the reference list. The texts were not identical but very similar, and the details they contributed to the picture of each study are convergent.

the first article (Tzohar-Rozen & Kramarski, 2013) compares an affective based self-regulation intervention group with a control group, and the second article (Tzohar-Rozen & Kramarski, 2017) compares an intervention based on metacognition to the same affective intervention (here called “meta-affect”) and control groups. However, this relationship is not explained in the articles themselves and could not be confirmed without contacting the authors. In both cases of presumed overlap, only one effect size is included in the meta-analysis and intervention details reported qualitatively are only reported once as discussed below. This review considered multiple publication types, and one fourth of the studies were dissertations (15). All these came from the US, and this is not surprising given that dissertations are likely to be released in the local academic language of the institution. This could imply some geographic imbalance in the “grey” literature included in this review. Types of included reports are shown by region in Table 12.

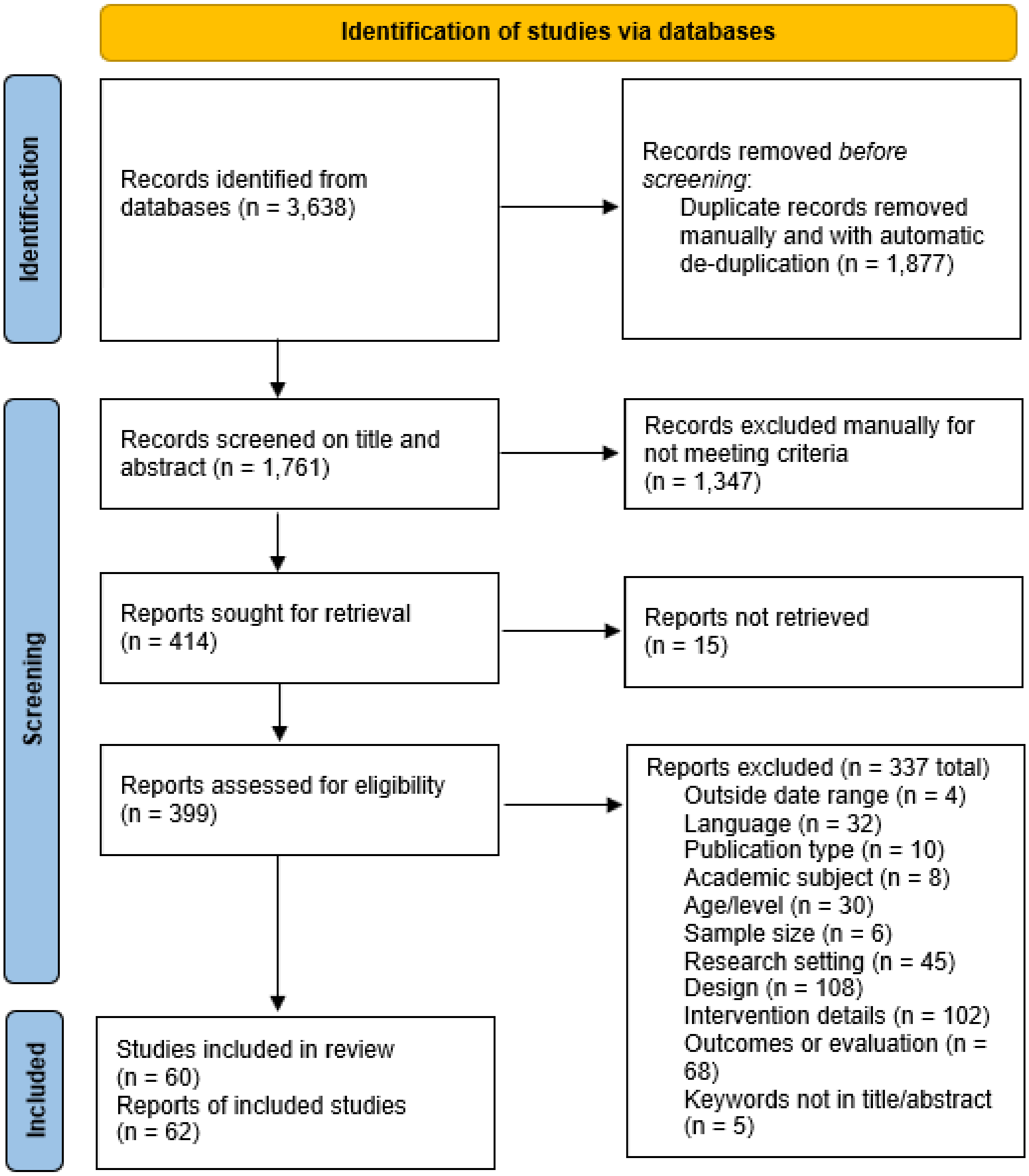

Identification of studies via databases
Identification
Screening
Included
Records identified from databases (n = 3,638)
Records removed *before screening*:
Duplicate records removed manually and with automatic de-duplication (n = 1,877)
Records screened on title and abstract (n = 1,761)
Records excluded manually for not meeting criteria (n = 1,347)
Reports sought for retrieval (n = 414)
Reports not retrieved (n = 15)
Reports assessed for eligibility (n = 399)
Reports excluded (n = 337 total)
Outside date range (n = 4)
Language (n = 32)
Publication type (n = 10)
Academic subject (n = 8)
Age/level (n = 30)
Sample size (n = 6)
Research setting (n = 45)
Design (n = 108)
Intervention details (n = 102)
Outcomes or evaluation (n = 68)
Keywords not in title/abstract (n = 5)
Studies included in review (n = 60)
Reports of included studies (n = 62)

Figure 5. PRISMA flowchart showing identified and included vs. excluded reports. Exclusions were coded with one or more reasons, based on pre-set criteria. Some reports were given multiple exclusion codes.

Region	Journal articles	Dissertations / Theses	Conference papers	Technical reports	Total reports
North America	10	15	1	0	26
Middle-East	16	0	2	0	18
Europe	10	0	0	1	11
East Asia	5	0	1	0	6
Oceania	1	0	0	0	1
All Regions	42	15	4	1	62

Table 12. Reports by publication type and region, from 62 included reports.

Included dissertations varied widely in length and quality but were generally more detailed than other report types, so including them could be instructive for users of this review. Four conference papers and one technical report were included in the review, while the rest (n=42) were journal articles. Of course, the use of electronic methods for searching and retrieving reports means that any not available online were missed in this review, and this potential limitation would be expected to affect especially those from the earlier publication years when paper-based publication was more common. In Figure 6, it can be seen that there is only one included study from each year between 2005 and 2007. It can also be seen that reports from the US and the Middle East appear most frequently across all years, those from Western Europe slightly less so, and those from East Asia, Eastern Europe, and Oceania are much less common.

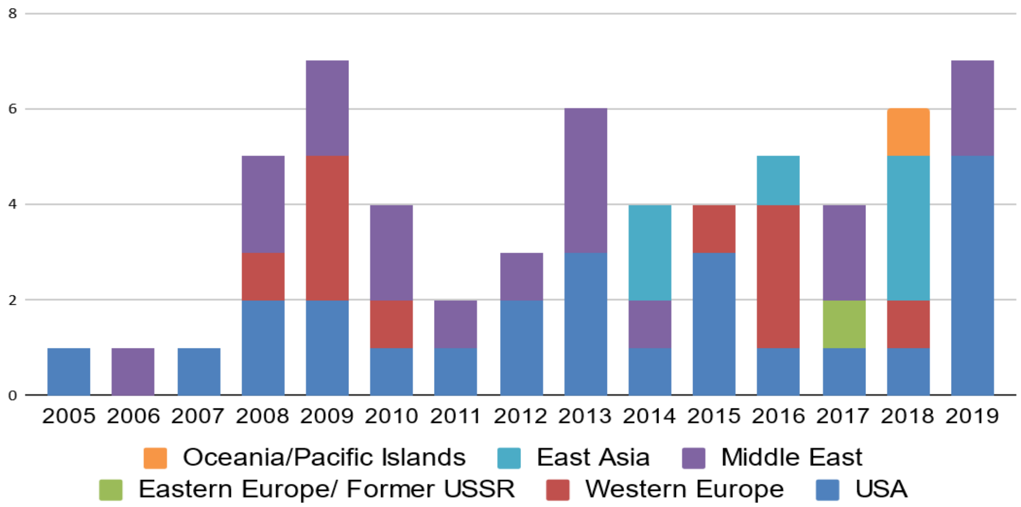


Figure 6. Included reports by publication year and region, with 62 total reports.

4.1.2 Geographic contexts

Figure 7 shows the geographic spread of the included studies, and it illustrates the level of international interest in raising mathematics attainment through MC/SRL based interventions. Note that in the later screening stages, studies from a diverse range of contexts had to be excluded because it was not possible to calculate an effect size or for design-related reasons. Thus, the reach of MC/SRL interventions is even broader than those represented here. For practical reasons, it was not possible to report on each study considered for inclusion and rejected. It is not surprising that the majority of final included studies originated in an English-speaking country since reports had to be available in English to be included, but it is notable that the US accounts for nearly half the set of studies (25), including all the dissertations (15), while the UK accounts for relatively few studies (2), and Canada, Australia, and New Zealand none. Israel accounts for the next largest set of studies (10), mainly those by Kramarski, Mevarech, and colleagues. Turkey contributed four studies to the final set, Germany and Indonesia[28] three studies each, Iran two, and the rest of the countries only one study each. Overall, North America, the Middle East, Western Europe, and East Asia were the most common regions for the research to be conducted, but there are caveats to this generalisation. By restricting the reports to English-only, to those with an explicit MC/SRL focus on mathematics, and to those with specific intervention criteria (e.g., not done at home), undoubtedly this review overlooks some research that could be illuminating. However, these

[28] Unfortunately, several additional reports from Indonesia were excluded from the final set because of language issues.

choices and others were made early in the process for theoretical and practical reasons, and they were applied to all studies regardless of the region, so in this way the resulting body of studies is “unbiased.” In fact, these criteria excluded some studies from well-known researchers in the field, while allowing some studies from lesser-known sources to be highlighted. Having such a large proportion of included studies originating in one country may lead to the impression that the review results are unbalanced, however, it should be remembered that the educational contexts in the US are extremely varied. While there may be commonalities between schools in different states, there are clear distinctions too. For this reason, it is important to consider other aspects of the research context, as is done below.

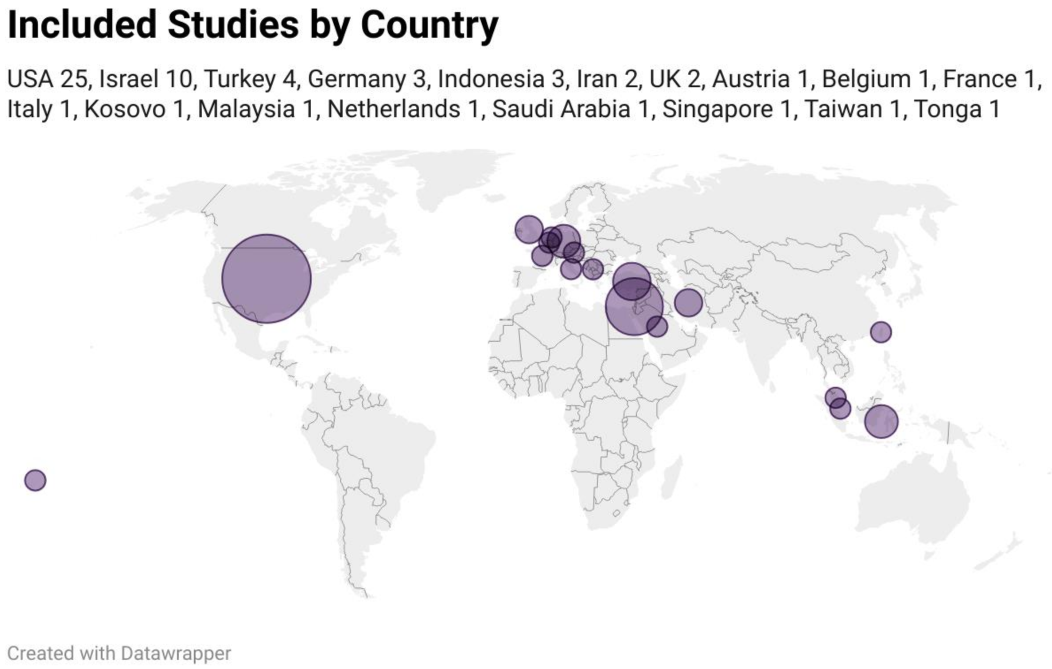

Included Studies by Country
USA 25, Israel 10, Turkey 4, Germany 3, Indonesia 3, Iran 2, UK 2, Austria 1, Belgium 1, France 1, Italy 1, Kosovo 1, Malaysia 1, Netherlands 1, Saudi Arabia 1, Singapore 1, Taiwan 1, Tonga 1
Created with Datawrapper

Figure 7. Included studies by country, with bubbles sized by number of studies.

4.1.3 Educational contexts

The intent was to capture as much information as possible about the educational contexts of the included studies, including the size and type of school, the type of community in which it was located, and the socio-cultural and linguistic background of the students involved. This was believed to be helpful information for stakeholders considering implementing interventions from the review in their own schools. However, it became clear that such information was often missing from included reports, especially those done outside the US. Dissertations/theses were the exceptions to this, possibly because their format allows for more in-depth considerations of context, yet the type of descriptions included are still variable.

Interpretation was needed to code contextual information. For example, if not directly stated, the type of community was inferred based on the name of the city or town, if reported. A code of "unclear" was used to mark whenever contextual information was absent, rather than assuming that participating students or schools reflected the "average" or "expected" socio-cultural reality for that country. For example, it was not assumed that all participating students for a study done in France would be ethnically French or speak French at home. Since reports of studies from the US were somewhat more likely to report contextual and demographic information, it may be that such information is considered more salient for educational outcomes within the US, but this idea was not evaluated in the present research, nor were subgroup analyses done based on contextual categories. The implication is that students from minority ethnicities or home-languages or who qualified for free or discounted school meals could be more at-risk academically, even if this was not stated outright. Several studies from the US and elsewhere reported this information without stating how it would affect the interpretation of the findings (e.g., Bond & Ellis, 2013; Cleary, Velardi, Schnaidman, 2017). See Table 13 for the frequencies of codes within each socio-cultural category. More than one code could be applied to a single study when the research was done with multiple sites or samples.

Table 13 shows that nearly half of studies (n=27) did not report community type (see also Figure 8), and the same number did not clearly report the type of school(s) (see Figure 9) involved in the research, and over half did not include information about students' ethnicities (n=38) or SES background (n=40) (see Figure 10). Home language is considered in this category as well as in the socio-cultural context because understanding linguistic expressions, not just numeric ones, is critical for mathematical functioning and development. Several primary studies mention language learner status as potentially adding to the challenge of learning mathematics. For example, Morales (2016), as a rationale for implementing a MC/SRL-based writing intervention in a mathematics class, states: "Most [English language learners] have difficulty with problem solving due to culturally-linked content and to vocabulary found in mathematics word problems" (p. 27). On the other hand, Wang et al. (2019, p. 3) interprets the academic "risk" of English learners in the US as being primarily linked to other factors, such as low income and parental education level. Only 12 studies out of 60 reported that some participating students had a different home language from that used in school, with nearly all of these being from the US (11 studies). Given the level of underreporting, it would not be possible to test or even theorise whether different intervention effects might be seen in different contexts, though again this might still form part of the stated rationale within individual studies.

Study Aspect	Code	All Studies	US Studies
Community Type	Urban	15	9
	Metropolitan	12	2
	Suburban	7	7
	Rural	2	1
	Unclear	27	9
Income or SES	Low or eligible for free school meals (FSM)	14	13
	Middle	7	5
	High	3	3
	Mixed SES	5	3
	Unclear	40	10
Ethnicity or Nationality	Specific ethnicity(s) mentioned	20	18
	Ethnically diverse	2	1
	Unclear ethnicity	38	6
Language	Home language different from school	12	11
School Type	Public	25	13
	Private	3	2
	Charter	2	2
	Head Start or other preschool centre	3	3
	Unclear or multiple types	27	5

Table 13. Numbers of codes applied for educational context for all studies and for those from the US. Note that individual reports could receive more than one code.

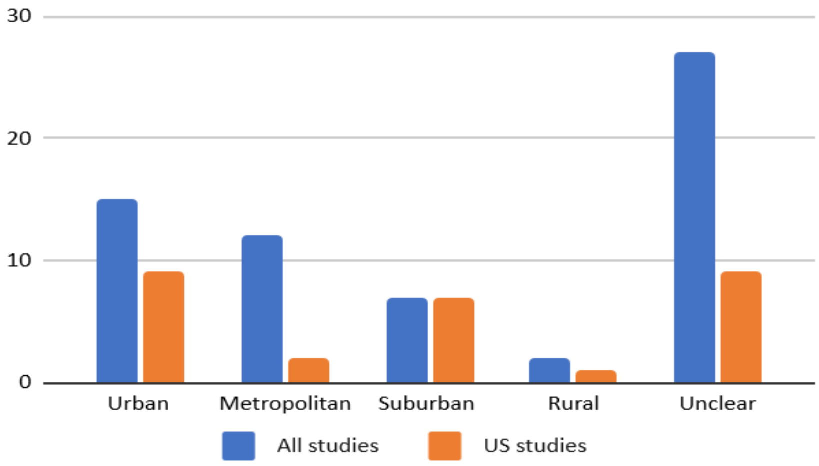


Figure 8. Community type codes applied to all studies and US studies. Some studies included more than one community type.

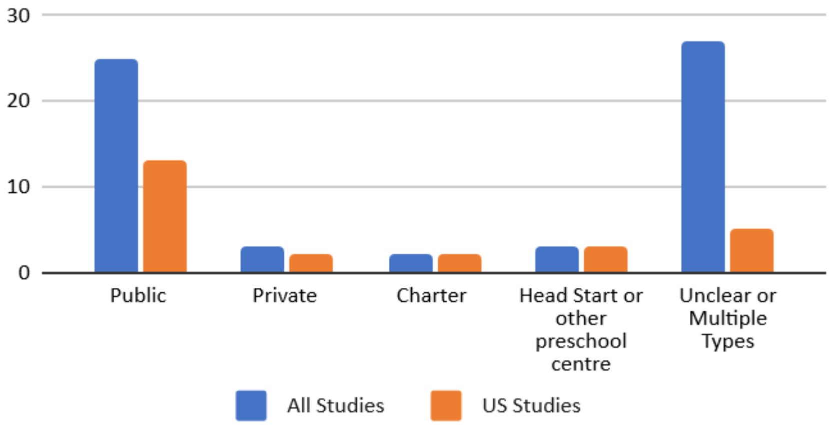


Figure 9. School type codes applied to all studies and US studies. Some studies included more than one school type.

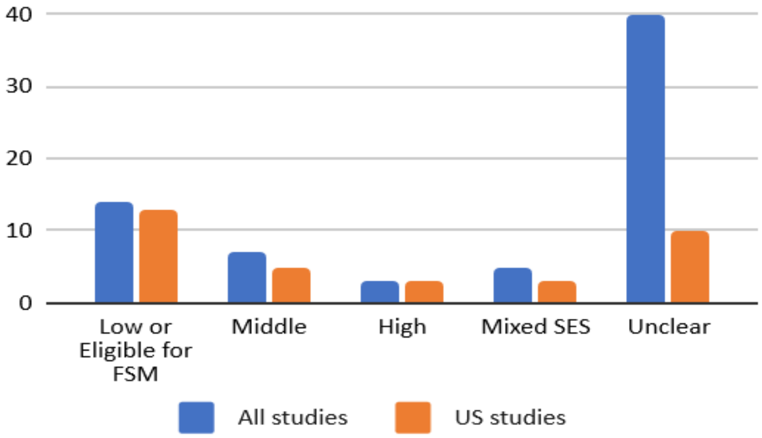


Figure 10. Socio-economic status (SES) codes applied to all studies and US studies. Some studies reported multiple SES backgrounds.

As covered in the theoretical background of metacognition and self-regulated learning, some interventions have been developed and tested specifically for use with learners having special needs, and this information was also extracted from the included studies where given (e.g., Edwards, 2008; Wang et al., 2019; Kang, 2010; Ford, 2018). Reported abilities and/or designations are shown in Figure 11, with US-based studies having clearer ability descriptions than other studies. However, as noted, all dissertations were from the US, and these also reported ability designations more clearly than other report types. “Unclear” was coded in 20% (n=3) of dissertations versus nearly 50% (n=19) of journal articles, for example. Only 10% (n=1) of US-based journal articles were coded as “unclear” about students’ abilities, which indicates such designations may be considered more relevant to the US context. To connect MC/SRL and ability, some researchers have proposed that deficits or higher than average levels of metacognition, self-regulation, or executive function may have a direct link to academic ability designations (e.g., Brown, 1977; Borokowski et al., 1989). Such theories predict students could benefit differentially from MC/SRL training based on their designations, but not all included studies referred to this, and over a third (24) of studies made no reference to ability or language learner status. Primary studies sometimes gave numbers or percentages of students with functional designations, whereas other reports simply indicated there were such students within the classes or groups allocated to different study conditions. Not having a goal of analysing effects based on ability-related subgroups, I only note whether students fitting different ability categories were present in the samples. It could be useful to undertake a further review of SRL/MC interventions specifically within the special-needs literature. Surprisingly, at least one study in the current review excluded students with disabilities from the analysis, even though such students were normally present in the class (e.g., Chen & Chiu, 2016, p. 272). In such cases, it would be helpful if research reports discussed how learners of various abilities were provided for in the local education system and under what conditions the target intervention might be expected to benefit them.

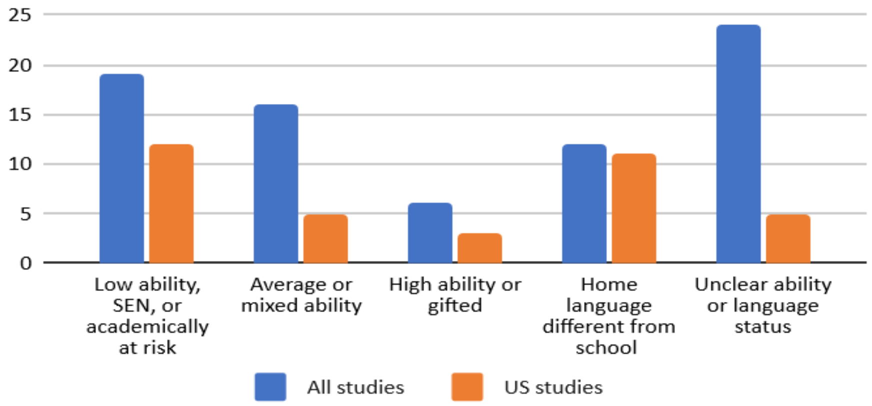


Figure 11. Ability and language status codes applied to all and US-based studies. Note that individual studies could receive more than one code.

4.1.4 School level, age, and gender

For the current review, only preschool and school-age pupils were included, and many studies with post-secondary or university students had to be screened out. Some research suggests MC/SRL interventions could have different effects for students of different age groups (e.g., Dignath & Büttner, 2008). For example, younger students may struggle with the domain knowledge and cognitive load required to successfully self-regulate. At the same time, they may experience more structured teaching and have less of a need or opportunity for MC/SRL. In the included studies of this review, there is a nearly even split between younger and older students. Five studies were done with preschool and kindergarten children and 27 involved primary/elementary students. For older pupils, there were 16 middle school, 11 high school or upper-secondary, and 3 unspecified secondary school codes applied. Some studies involved pupils at multiple school levels. In the current review, the higher number of studies done with primary and middle school pupils may be related to several issues or intentions in addition to a desire to demonstrate effectiveness of the interventions in younger learners. First, the inclusion criteria may have affected the distribution of school levels included. I considered interventions to be most applicable for classroom teachers when they did not require extra time outside of the normal school day, thus after-school programmes or those relying heavily on homework to reinforce the MC/SRL training were excluded. This means included studies, in some cases, would have needed to reallocate teaching times and classrooms to deliver the MC/SRL interventions, especially when these were done on an individual or small group basis, and this presumably is more possible in the earlier grades. Children in primary classrooms often have one teacher for multiple subjects, and it makes sense that there would be greater scheduling flexibility. It should be stated here that reports sometimes mentioned, but often did not, whether the intervention was done completely during the normal mathematics lesson or otherwise, and it is possible that some included studies were actually delivered outside of the normal day without this being clear in the reports.

The second reason included studies may have focused on primary and middle school students is that mathematics lessons in these grades are more likely to involve a range of ability levels. Interventions designed to assist learners in being more strategic may be most appealing for teaching mixed ability groups, rather than older, more homogenous groups. Finally, the MC/SRL interventions for younger ages in this review frequently featured games and opportunities for creative expression (Pappas Schattman, 2005; Wang et al. 2019; McClelland et al., 2019), all designed to encourage active engagement with the mathematics curriculum. Such activities may be seen as less important for older pupils, who may be expected to be more achievement-minded. In Figure 12, school-level codes applied to included reports are shown. Codes were applied based on the study descriptions, with no attempt to adjust for international differences in age or curriculum at each grade or year of school. Less commonly, school level was inferred based on the reported age of participating students. While the US accounted for all of the preschool studies, in general US-based studies were somewhat more likely to be done with older pupils. Regarding preschool studies, attending preschool is non-compulsory for US children but has been promoted as a measure to overcome pre-existing disadvantages. Local and national programmes, such as Head Start, offer preschool funding for low-income families, and several included studies were done in such contexts (Schmitt et al., 2015; McClelland et al., 2019) Some research has demonstrated mixed results of preschool for disadvantaged students (Loeb et al., 2007) most notably in the area of social skills and behaviour, making these included studies even more relevant.

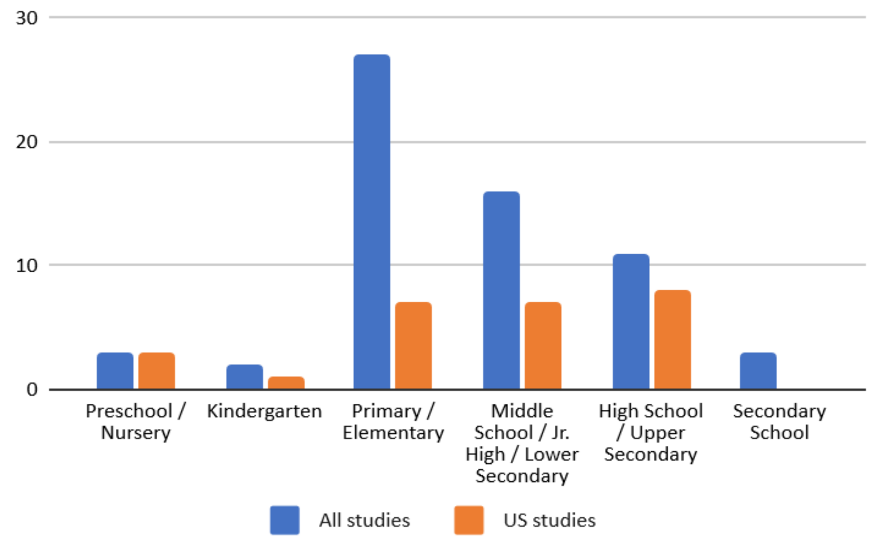


Figure 12. School-level codes applied to all included studies and US studies.
Individual studies could receive multiple codes in some instances. "Secondary school" was used when it was not clear whether the study was with lower or upper secondary students.

Figure 13 shows the spread of average participant age in included studies, with most studies concentrating on ages eight to 14, roughly equivalent to upper-primary and middle-school students. Twenty-four studies did not report age clearly, in which case average age was estimated based on the year-level. Because most studies did not report individual participants' ages, an estimation procedure was used. Each study was only coded for one average age even when there were multiple ages or year groups, and the bars show average age up to the next integer. An average age of 4.5 would be coded as 4, for example. Because of this averaging, this chart does not reflect all ages included in the studies. The "Red Light, Purple Light" studies (i.e., those by Tominey, Schmitt, McClelland and colleagues) were done with ages three to five but were coded with an average age of four. Pupil age is not a main focus of this review, but it could be a factor affecting intervention effectiveness. Even within a single grade there may be differential outcomes for pupils at the high and low ends of the age spectrum (Dhuey et al., 2019). The fact that so many reports included here failed to capture this information, while every report included at least the general school level and usually the grade, is concerning. It suggests that inherent developmental processes could be obscured in how progress in mathematics learning is represented within the included studies.

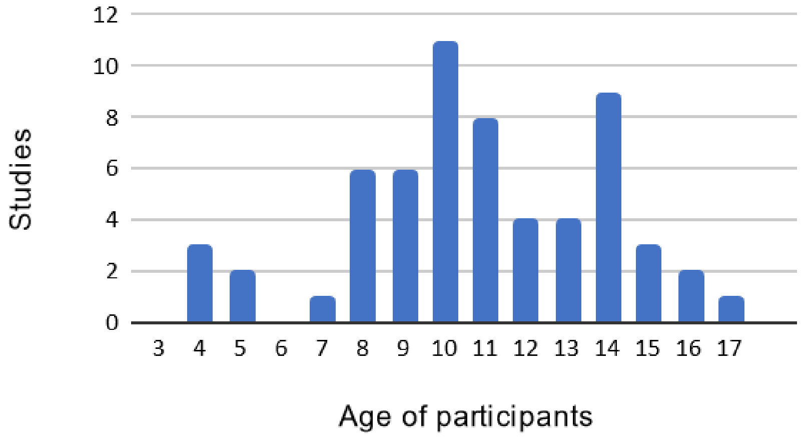


Figure 13. Average age of included participants in years by study and numbers of studies.

Based on previous studies of MC/SRL-based interventions that reported different effects in females and males (e.g., Cardelle-Elewar, 1990), I also coded the gender of participants, though this information was missing from eight reports. The great majority of student groupings were mixed-gender (50 reports), with four reports coded as having female-only groups and four reports with male-only groups. Fully gender-segregated classes were mainly found in studies from the Middle East (Abdolhossini, 2012; Babakhani, 2011; Mevarech & Amrany, 2008; Rizk, Attia, & Al-Jundi, 2017), but in one case a school in Austria had some mixed (low numbers of females) and some gender-segregated (male-only) classes (Fößl, et al. 2016). In no reports was a rationale given as to why single- or mixed-gender classes were used in the schools or in the study. No reports mentioned including genders or identities other than male and female. Reports sometimes mentioned the percentages of each gender in each treatment group, or in the overall sample, or they simply indicated that gender levels were approximately equivalent in each group without giving exact numbers. Reports rarely told how they captured gender information about participants. Due to the prevalence of missing information, it is not possible in this review to consider differential intervention effectiveness based on the gender concentrations in the studies, but several studies found different outcomes by gender (e.g., Edwards, 2008; Falco, 2008; Jackson Jackson, 2012).

4.1.5 MC/SRL Interventions and comparison conditions

This subsection describes important aspects of the study conditions in the included reports, starting with an introduction to the intervention and comparison groups, moving to the intervention details reported, the timeline or "dose" of the interventions, the roles of intervention leaders and classroom teachers, training for and fidelity of implementation of the interventions, participants' views of the interventions, the mathematics content included in the intervention period, and finally the MC/SRL stages, activities, and strategies focused on in the interventions. The section ends with a brief look at the theoretical foundations of the included interventions and a summary of the key findings about the interventions.

Various designs were included in the review, but studies needed at least one intervention group based on MC/SRL training for students and one comparison group not based on training in MC/SRL. Thus, comparison groups could have business-as-usual (BAU) or active controls, and sometimes included both. Active controls could share common elements with the experimental groups, such as the mathematical tasks, the general class structure and materials, and, in some cases, the same teachers. The main requirement was that active controls could not include MC/SRL training for students. Sometimes studies were excluded from the review when controls were presented as not being based on MC/SRL but they still featured what I interpreted as MC/SRL training, such as in calibration-type exercises (DiGiacomo & Chen, 2016) and metacognitive practice in a different academic subject (Wessman Huber, 2010). The reasoning was that these elements would make the comparison group too similar to the interventions in other studies where the same elements are used in MC/SRL training. For example, in Wang et al. (2019), the active control group ("base condition," p. 341) used a rewards-based motivational system and a structured problem-solving approach based on schema-based instruction, which are elements used in MC/SRL interventions from other studies in this review. In this case, only the active control group was excluded and the SRL intervention and BAU comparison group were retained.

The choice was made to combine all MC/SRL intervention groups for the purposes of calculating an effect size based on the review goal of comparing MC/SRL and non-MC/SRL conditions. There was no intention to compare different types or amounts of MC/SRL training within the same study in the summary estimate of effect. The qualitative analysis therefore also does not distinguish between different MC/SRL conditions within the same study; however, the original groupings are given in Appendix 9, which shows the study designs and describes the interventions. The table in Appendix 9 also shows whether studies used BAU or active controls, or both, in which case these have also been combined in the meta-analysis so that a single effect size was generated for each study. This avoids the problem of double-counting the control groups sometimes seen in the original reports, where multiple intervention groups were compared with the same control (e.g., Cross, 2009; Kramarski & Dudai, 2009).

Study groups	**1 Comparison condition**	**2 Comparison conditions**
1 MC/SRL condition	47	5
2 MC/SRL conditions	5	1
3 MC/SRL conditions	2	0

Table 14. Included studies by number of MC/SRL conditions and comparison conditions.

Studies using two or more comparison groups (see Table 14) were presumably trying to control for other components of the intervention besides the “active” MC/SRL elements. For example, Arroyo and colleagues (2007) tested a tutoring software package that provided regular metacognitive feedback and prompts to users (intervention) or allowed users to access non-metacognitive hints on-demand (“tutor control”). They also included a BAU group with normal teaching and no access to the programme (“no tutor control”). In this case, it is apparent that the researchers wanted to differentiate the effects of the MC elements from those of the software itself. Barrus (2013) used a similar approach, comparing a BAU control, a comparison group which practised with the ALEKS® programme, and an intervention group that completed computer-based SRL training modules developed by the researcher in addition to ALEKS®. Shamir and Lifshitz considered the effects of using an e-book with metacognitive guidance (intervention) or without it (active control), and they also included a BAU control not using the e-book. Bond and Ellis (2013) tested the effects of a newly-made mathematics teaching unit that closed lessons with a reflective activity (MC/SRL intervention) or a non-reflective review (active control). They also included a BAU control using another unit of curriculum. Jackson Jackson (2012) examined four study conditions in a factorial design: high and low “communal learning” and with and without an SRL component.

There were also studies that considered different versions of the MC/SRL training programmes. As stated, it was outside the scope of the review to compare effects based on different forms or "dosages" of MC/SRL training. However, the original included studies may have had such a goal or may have planned to test the combined impact of specific MC/SRL elements. For example, Dresel and Haugwitz (2008) featured several study groups. In one group, students practised problem-solving with a computer-based intervention, MatheWarp, which offered them "attributional feedback enriched with metacognitive control questions" (p. 7). There was also a "placebo condition," as well as a group that used the programme without the metacognitive questions, but with the attributional feedback. I considered the latter group an intervention rather than a comparison group because attributions have been seen as part of the MC/SRL process (e.g., Bandura, 1991), but it is not clear what the authors' original classification was. In Kramarski and Friedman (2014), pupils in all study conditions worked with a learning software package with metacognitive prompts that were either automatic or chosen by the student, or with no metacognitive prompts (control group). In another study (Kramarski & Zoldan, 2008), metacognitive questions (i.e., IMPROVE) are used alone or in combination with an error-diagnosis approach, and the latter is also used by itself in one condition. Because considering the source of errors has been used as a stand-alone metacognitive approach (cf. Heemsoth & Heinze, 2016), all three conditions were coded as MC/SRL interventions, and there is also a BAU control. Kramarski and Dudai (2009) used two different IMPROVE conditions, one which trained students to evaluate their own solutions to a problem ("self-explanation guidance," p. 385) and one which trained them to provide feedback on peers' solutions ("group feedback guidance"). In the control group, students could work with peers on the same tasks and had previously received training in mathematical explanations, but they did not use structured metacognitive questions to guide and check their solutions. In Tzohar-Rozen and Kramarski (2017), as stated above, the two interventions differed by being based on either emotional regulation ("meta-affect") or cognitive regulation (IMPROVE). Cross (2009) used three intervention groups training students in verbal or written mathematical argumentation, or a combined approach, as compared with a teacher-centred BAU control. Finally, McClelland and colleagues (2019) implemented a previously tried intervention ("Red Light, Purple Light Circle-Time Games") based on behavioural regulation and added a condition that focused explicitly on early mathematics and literacy development (SR+).

4.1.6 Reporting of MC/SRL interventions

Review reports were next coded for the level of detail used in describing study conditions. Results of this coding are shown in Table 15. Reports coded with "high intervention detail" included enough description, such as samples of materials and schedules of activities, for all or much of the intervention to be replicated or reconstructed in a new teaching context based on the report alone. Those coded as "moderate detail" would permit users to reconstruct some intervention components, while those coded as "low detail" would require consulting the original authors or other sources to be reconstructed. Level of detail was assigned holistically rather than being based on the length of the text because interventions differed greatly in their complexity. Some interventions could be fully described in a few sentences and fully implemented in a few minutes per class session, while others would require the use of multiple worksheets, teaching scripts, and other materials to implement. Where reports referred to other sources for additional descriptions of interventions, this was noted in the coding, but the other reports were not added to the review based on the decision not to use "snowball" or other methods that could be unreliable. If external sources of intervention detail had been admissible during the screening stage, this would have changed the nature and practical requirements of the review. Likewise, the data extraction here considers only those reports found in the systematic searches and retained through the screening process. Whenever there was an appendix or online supplemental material available from the same source as the main report, this was considered another section of the main report, and the inclusion of such generally resulted in a higher level of intervention detail. However, in the case of Baliram and Ellis (2019), the supplemental file was a condensed version of the article geared toward practitioners that did not add intervention detail.

Intervention detail	All included reports	Samples of activities or materials	Schedules of activities within or across sessions	Samples of students' work with the intervention or dialogue	Samples of intervention leaders' work or dialogue	Intervention described elsewhere
High	16	14	14	6	3	3
Moderate	41	24	19	6	1	10
Low	5	2	1	0	0	1
Totals	62	40	32	12	4	14

Table 15. Intervention detail codes applied to 62 included reports of 60 studies.

From 62 reports of the 60 included studies, 16 reports were coded as “high detail,” 41 as “moderate,” and 5 as “low.” This is partly a result of the review protocol, since reports were excluded that did not include enough comprehensible description to determine the nature of the intervention. Three reports were excluded during the data extraction phase because there was too much uncertainty about the intervention due to low description. Of those retained but coded as “low detail,” Schmitt (2013) described one activity, the “Red Light, Purple Light” game, and referred to other reports of the intervention, and Tzohar-Rozen and Kramarski (2017) included the self-questions used in the metacognitive and meta-affective training conditions. Three other “low detail” reports described interventions based on modelling a mathematical task visually (“VStops,” Abdullah, Halim & Zakaria, 2014), using classroom dialogue, and metacognitive journaling (Aminah et al., 2018), or reporting broad MC/SRL objectives for each lesson (Abdolhossini, 2012) without providing samples of teaching schedules or activities.

Reports described as “high detail” include nine out of 42 journal articles and seven out of 15 dissertations. The length of a dissertation may admit greater intervention reporting, although this was not always used to advantage, but online journals now frequently feature supplemental materials, as noted above. All conference papers (3) and technical reports (1) were coded as “moderate detail.” Overall, two-thirds of reports (40) included samples of activities or materials used in the interventions, such as worksheets, screenshots from software packages, mathematical tasks used in learning activities, or self-questioning guides. Just over a half of reports (34) included a schedule of activities within a sample lesson or multiple lessons, while less than a quarter of them (12) included samples of students’ responses in the intervention, such as completed tasks or dialogue transcripts. Even fewer reports (4) included the responses of intervention leaders, such as written or verbal instructions or feedback on students’ work.[29] Thus, intervention materials provided in reports represent what was intended more than what was actually done in the sessions. Reports frequently, though not always, discussed how intervention leaders were trained and whether attempts were made to ensure intervention “fidelity,” though most did not report fidelity levels quantitatively.

The decision to present the “as intended” version of the intervention in a report rather than that actually used in the classroom could reflect the limitations of data-collection within the research site. It could also be seen as aligning with the principle of “intention-to-treat” (Gorard, 2013, p. 168) in which study participants are assessed based on their assigned study condition, even if they missed sessions or switched to a different group. However, underreporting the

[29] Teaching scripts, like that used in Sings Jenkins (2009) were not considered teachers’ authentic responses but were instead coded under “samples of activities or materials.”

study conditions as implemented could lead to underestimating the intervention effect or overlooking key active elements. As a hypothetical, if teachers are trained to use a journaling intervention, and it is overlooked that they assign extrinsic rewards based on the completeness of pupils' entries, then the interpretation of the study results could be skewed and difficult to replicate. Within the current review, many studies expressed a social-cognitive orientation (e.g., Barrus, 2013; Kang, 2010; Sings Jenkins, 2009), in which it is seen as important that students explain their mathematical reasoning in classroom dialogue and prompt each other to think more deeply, with teachers scaffolding the process of making thinking explicit. Students' internalisation of classroom dialogue can then become a catalyst for greater MC/SRL skilfulness. A failure to report students' responses during the intervention may belie the social-cognitive explanation or suggest there are other more important mechanisms of effectiveness at work. It should be noted that in many cases, studies did use surveys and other assessment tools to check for changes in students MC/SRL processes, but it is outside the scope of this review to synthesise these results, and this is not a substitute for seeing the intervention in action. Teachers could especially benefit from samples of authentic intervention responses, especially since they enable a clearer operationalisation of core concepts and show the practical ways the intervention may differ from the standard approaches.

4.1.7 Study timeline and intervention "dose"

This review also examined the session length, frequency, and overall timeline of the MC/SRL interventions. While most studies (n=49) reported intervention timing, it was in some cases unclear or incomplete, which made comparing the intervention "dose" between studies a challenge. Reports used different and sometimes non-comparable units of time, such as "class hours" (Ubuz & Erdoğan, 2019, p. 136) and it was often unclear whether the intervention encompassed a whole class period or only part. In some cases, descriptions were inconsistent throughout the report (e.g., Babakhani, 2011, p. 566). Thus, interpretation and inference were required to determine the intervention dose, and this could limit the trends seen. As shown in Figure 14, most studies were for 12 weeks or shorter (n=50), with 23 studies indicating the intervention was delivered for a month or less, and 27 reporting an intervention period of between five and 12 weeks. Six studies ran their interventions for between 13 weeks and the length of one semester, and four were between a semester and a full academic year. Note that sometimes the length of time given also included assessment activities and it was not always possible to isolate just the intervention period. The basis for choosing a certain length for the intervention was rarely discussed and could have been related to practical considerations as much as the theoretical rationale.

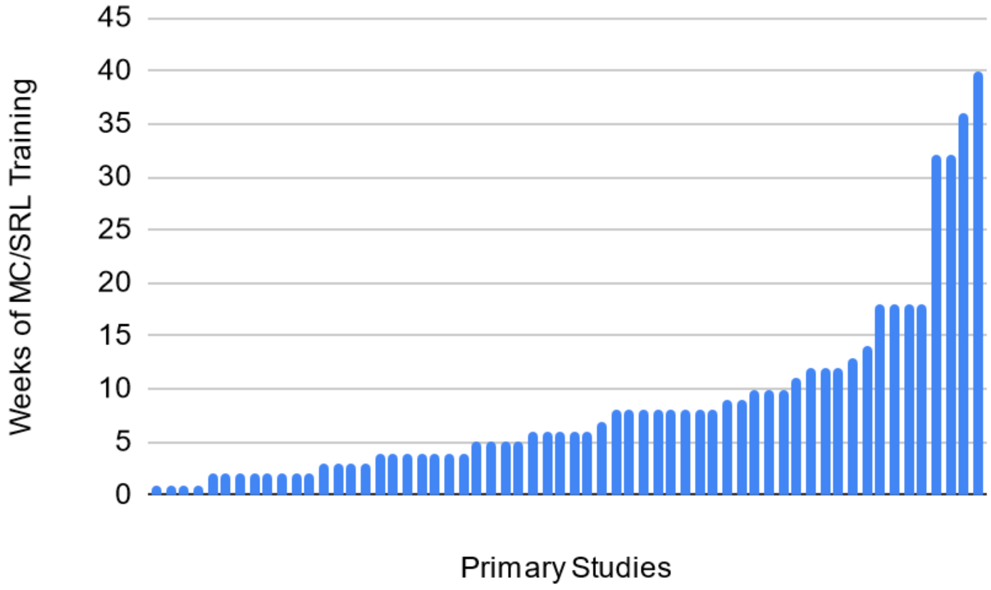


Figure 14. Length of intervention in weeks for included studies.

In terms of treatment frequency, most interventions were offered either daily or almost daily (16 studies), or two to three times per week (20 studies) during the intervention period featuring sessions between 30 minutes and an hour in length. Nine studies employed a weekly intervention schedule, while three studies featured bi-weekly or monthly interventions. For 12 studies, the frequency of the intervention was not clearly reported. It was only possible to extract data about the length of each intervention session for 42 studies, with 18 being unclear. Of those reporting length of the sessions, about 1 hour was the most frequent (20 studies), with 15 to 30 minutes being the second most frequent code (15 studies). In five studies, the intervention was up to 15 minutes per session, one report stated the intervention was 80 minutes long (Lestari & Jailani, 2018), and one had sessions that were 4 hours in length (Kramarski & Friedman, 2014). If multiple session lengths were reported, the longer value was coded. In some cases, the length may refer to an entire mathematics teaching session in which explicit MC/SRL activities only occupied a portion of the time, with other activities being shared with the comparison group. It was not always possible to distinguish time spent on activities shared with the comparison group from that spent on unique features of the intervention. Frequently, the MC/SRL training consisted of a few elements used at different points throughout a session, such as metacognitive prompts for structured problem-solving. In these cases, it seems reasonable to consider the entire session as contributing to the MC/SRL training. Number of intervention sessions was also coded, though this was not reported for 14 studies. The range was between three and 56 sessions. The most common was six to 10 intervention sessions total (16 studies), while the second most common range was between 16 and 20 sessions (10 studies). Seven studies used between three and five sessions, while six studies offered 11 to 15 sessions. Only seven studies offered 21 sessions or more. Within studies that reported the total length of the intervention as well as the number of sessions (n=46), most studies (n=37) were up to 12 weeks and 20 sessions. Within this group, six studies had up to five sessions spread over up to three weeks. There were 11 studies that included six to 10 sessions spread over six weeks or less, and there were 13 studies that implemented 11 to 20 sessions over a period of four to eight weeks. Outside of these concentrations, there were no clear patterns in the intensity of the intervention meetings. The average length of intervention sessions makes sense considering they were done at school, mostly during regular mathematics classes. Overall, it seems that most interventions were done for at least 30 mins on multiple days each week, for a period of one to three months. This is important because it shows a commitment to sustained MC/SRL training and the need to give classroom communities time to adopt new discourse norms and patterns of thinking.

There were certainly exceptions to the “dose” patterns described above. For example, O’Neal (2016) implemented very short, writing-based MC/SRL sessions in an Advanced Placement (AP) calculus course. Pupils used three types of prompts to complete daily reflective journal entries about their learning during the last 5 minutes of 29 class periods. The implementing teacher encouraged students to elaborate their responses, after they were initially too limited (pp. 41-42), but this could have been a challenge given the very limited time allocated. With somewhat longer sessions, Kang (2010) implemented 15-minute goal-setting activities within daily mathematics lessons for students with special needs, but the total intervention period was only 1.5 weeks (seven sessions total).[30] Byrd (2019) used a student response system (SRS) known as Classflow™ to task students with responding to learning items anonymously and then giving “elaborative” feedback on the responses of others. This process was intended to stimulate students to explain and justify their thinking and share

[30] Kang (2010) included two studies, but only one is included in the review because of design issues. In Study 1, the goal-setting intervention was used with a wait-list control group. After all the students had received the intervention, they were re-randomised for Study 2, which used reflection rather than goal-setting activities. It was decided that Study 2 should be excluded based on potential confounds from Study 1, since the same participants are used.

productive learning strategies (p. 52). The teacher also provided the treatment group with explicit instruction in goal-setting, monitoring, and reflection strategies, but in total the interventions were done for 10 minutes, three times weekly for 12 weeks. While practicality constraints in these studies may have limited the time researchers had to work with students, these examples at least illustrate the possibility of delivering MC/SRL training in “bite-sized” rather than intensive sessions. They might especially appeal to teachers looking to begin using discrete MC/SRL elements in their instruction but without the time and other resources to use most of the interventions from this review. Alternatively, smaller-scope interventions could be used in combination. Teachers could use an SRS approach to encourage dialogue, then have students complete reflective journals or set learning goals following this dialogue. As mentioned, Byrd (2019) used multiple MC/SRL elements even with shorter sessions (10 minutes), and previous reviews have presented different perspectives on the complexity of MC/SRL interventions. This topic is explored in more detail below as the specific MC/SRL elements are considered for the present review.

At the other end of the “dose” spectrum, Motteram et al. (2016) reported on the large-scale evaluation of ReflectED in the UK, using 28 structured, 30-minute lessons about learning strategies, and asking students to reflect on their own learning activities at least once weekly throughout the year. This resulted in 56 coded MC/SRL sessions in the review, but there was likely great variability in the actual numbers, given that the study was implemented across thirty schools and 70 classes. Notably, mathematics achievement was a pre-planned outcome of the study, but the intervention was not limited to or specifically designed for use in mathematics classes. It is not clear how often the intervention was used in mathematics compared with other classes, thus, the idea of “dose” here could be interpreted differently. Sarette (2014) was a smaller-scale study in a single school that combined training in behavioural self-regulation with goal-setting activities and discussion to help students discover and implement learning strategies. The training was delivered in 55-minute sessions twice a week over most of the school year, with the researcher initially modelling the approach in class with a gradual transition to delivery by the regular classroom teacher. Importantly, the MC/SRL activities were tailored to students’ current needs through an ongoing collaboration between the researcher and the teacher, but structured lesson-plans were still used. In Barrus (2013), students in the intervention group worked through 19 researcher-designed, computerised self-regulation training modules, used in tandem with the ALEKS tutoring software. The MC/SRL modules were self-paced and completed in daily, 20-minute sessions for 12 weeks during one semester, while comparison group students either received BAU mathematics teaching or used ALEKS without explicit MC/SRL training. All these examples illustrate the range of intensity of MC/SRL interventions used in schools.

4.1.8 Intervention leaders and classroom teachers' roles

Next, the role of the intervention leader needs to be considered. All the included studies were done at school, as studies reporting interventions outside the normal school day or relying on homework activities were disallowed in the original review criteria.[31] The intention was to estimate the effects of classroom-based interventions, but not all included studies were conducted by the normal classroom teacher. As shown in Table 16, 17 out of 60 studies included researcher-led MC/SRL sessions, while regular classroom teachers were leaders in 37 studies. Interventions were led by other school staff in four studies. For example, Cleary, Velardi, & Schnaidman (2017) trained "an assistant principal, two counsellors, and a school psychologist" (p. 35) to deliver the intervention in small groups of pupils. Eight studies were based on technology use, such as computerised mathematics tutoring. In four studies, MC/SRL training was conducted by researchers who were also teachers or other school-staff members. In three cases, the intervention leader was unclear. Sometimes there was shared delivery of the intervention. In seven studies, the classroom teacher and a researcher or other school staff member led the intervention together. In six studies, the classroom teacher or researcher led a discrete part of the MC/SRL training, while students received computerised training for the rest of the sessions. Overall, 26 studies were led by classroom teachers alone, while eight studies were led by researchers alone. In general, the review shows a focus on ecological validity within these studies, with the research designed to be directly applicable to teachers' practices. It is anticipated that when the interventions were led by researchers there would not be the same impact on teachers' beliefs, their feedback to students, and their support of metacognitive dialogue.

Intervention leader	**Regular classroom teacher**	**Other School staff**	**Researcher**	**Technology based**	**Researcher-as-teacher or other school staff member**	**Unclear**
Studies	37	4	17	8	4	3

Table 16. Codes applied to studies for intervention leaders.
Note that the codes exceed the number of studies because some studies included multiple intervention leaders.

[31] Because normal mathematics teaching often includes homework, studies were only excluded if the homework was different for the intervention group and was an important part of the MC/SRL training.

Information about how teachers or other leaders were trained and supervised in the intervention was also extracted from the included reports and is shown in Table 17. Most often (32 studies), researchers distributed intervention manuals, lesson plans, or other written guides, and they frequently referred to these as encouraging greater fidelity of implementation. Nearly as often (28 studies), live training sessions were led by researchers or other intervention developers. In 22 studies, both live sessions and written materials were provided. In two studies, video-based training was implemented, and in ten studies there were classroom visits, emails exchanged, or other forms of supervision or collaboration between the researchers and programme implementers. In 17 cases, the intervention training was not clearly reported. For those that did report training, use of written guides could make the implementation more consistent and replicable, while holding live sessions could facilitate clarification of expectations and a deeper engagement with the theory of the intervention. However, no studies compared different training modes side-by-side, and they could have complementary benefits.

Training mode	**Live sessions**	**Written guides**	**Videos**	**Supervision or collaboration**	**Unclear**
Number of studies	28	32	2	10	17

Table 17. Codes applied to studies for intervention training.
Each study could receive multiple codes.

It was also noted where researchers checked if the intervention was implemented faithfully, though some reports did not use the term “fidelity” for such checks. Twenty-five out of 60 studies included a description that was coded as a fidelity check, while in 35 studies it was not reported that this was done, or fidelity was unclear. Where checked, only 13 studies stated that fidelity was high, nine did not report a fidelity level, and three studies indicated there were problems with the implementation. For example, Ford (2018) observed every other intervention session led by the classroom teacher and “document[ed] the presence of the core features of metacognitive training . . . The same checklist was used in the control conditions to evaluate program differentiation and determine whether the problem-solving strategies group was provided any key elements of the metacognitive training” (p. 52). Through these structured observations, the researcher identified key elements of the intervention that were not being implemented but noted improvement following feedback and “retraining” (p. 54) of the teacher. Sings Jenkins (2009) reports that one teacher was unable to implement the intervention as planned and left the study, and Wijaya et al. (2018) indicates that teachers sometimes became uncertain or impatient when students did not respond quickly or in predicted ways (p. 9). The latter study focused on the concept of “opportunity-to-learn,” and trained students to approach authentic tasks critically to determine, for example, what information is relevant for solving them. They were also expected to justify their answers. That teachers had difficulty supporting productive dialogue suggests a need for more careful training around the theory of the intervention and how core concepts are operationalised. Collingwood & Dewey (2018) reported high fidelity overall, but there was some variability between the different teaching assistants who led the intervention. They also stated that some of the intervention elements were less consistently implemented than others, such as teachers’ responses to boost learners’ mathematics self-concept (p. 84).

As indicated above, actual implementation fidelity was missing or unclear in most studies (n=44), even nearly half of those that stated implementation was monitored (9 studies). For example, Shilo and Kramarski (2019) analyse classroom discourse from video-recorded intervention sessions, highlighting differences between the study groups. They also state that structured fidelity checks were used to give feedback to the intervention leaders, but they do not report the substance of the feedback or give an overall fidelity level (p. 630). Bond and Ellis (2013) similarly state: “The researcher closely monitored progress throughout the investigation to ensure that lesson scripts were followed, confidentiality was maintained, and disruptions were avoided” (p. 230). This implies, but does demonstrate, that implementation fidelity was high, and it potentially obscures actions taken to enforce it during the intervention period. Another way to consider fidelity of implementation would be to ask how the intervention training may have impacted other, unintended aspects of the classroom teaching. Within the review, there is almost no mention of such potential spill-overs in teachers’ practice, and it was not possible to code this comprehensively. However, above examples illustrate teachers sometimes had trouble re-adjusting classroom norms to suit the MC/SRL training. Assuming they are successful in this, it might be worth examining other explicit or implicit changes to the classroom that follow. From an evidentiary perspective, high fidelity might not guarantee the intervention has *caused* any outcomes observed in a study, since a causal argument relies on numerous factors, but low fidelity would certainly call this into question.

4.1.9 Views of the intervention

An issue connected to fidelity of implementation is the social acceptability of the intervention. Studies frequently assessed students' MC/SRL skills and attempted to conclude whether these had been impacted by the training. Less often, reports included samples of students' dialogue or other "snapshots" of how they responded to the intervention. However, students' subjective perceptions of the intervention could also be relevant in pedagogical decision making, and this was coded where surveys or interviews with students on this topic were reported, or where researchers made observations about students' perceptions of the intervention. As shown in Table 18, only positive responses from pupils are noted in 10 reports, while eight report both positive and negative responses, and in one report only negative responses are reported. Forty-three out of 62 reports did not include students' views of the intervention. Where positive responses are indicated, students' verbal or written quotes frequently illustrate presumed mechanisms of effectiveness for the MC/SRL interventions, such as increased mathematics related self-concept and better control and monitoring of behaviour and task performance. For example, Mandaci Şahin and Kendir (2013) quote one student, Yeter:

> Mathematics has started to be more entertaining. I have started to like thinking about the problem, planning and drawing figures. I used to make a lot of mistakes while solving a problem, for the process was too fast. Our teacher asked us whether we had understood or not. When we reported that we had not understood, he/she would go over the problem again too quickly. This time I could not report that I had still not understood. With this method, we solve problems slowly through games. I do not make mistakes now. Even if I make a mistake, I see where I am wrong and understand problems in a better way. (p. 1787)

As the above example shows, pupils also mentioned increased enjoyment of mathematics learning activities and pointed out specific aspects of the MC/SRL approaches they found helpful. As noted above, negative responses were reported less often and showed students sometimes found the MC/SRL-based activities difficult, boring, or unhelpful. For example, Lee, Yeo, and Hong (2014) trained students in a structured approach to problem-solving based on detailed task-analysis and diagramming, which seemed at times unwarranted for easier problems or took too long to utilise during an assessment (p. 473). The researchers concluded that students need to be explicitly taught the potential benefits of such an approach.

There are other included examples that show students could experience negative affect related to MC/SRL interventions. Tominey and McClelland (2011), which used the "Red Light, Purple Light" intervention, found that some children were passive or apprehensive about participating in the circle games, but their response improved when given the chance to lead the games themselves (p. 513). In Motteram et al. (2016), pupils used iPads to record video, audio, and still images, and to annotate these samples of their work with reflections on learning. These materials were made available to others through the "Evernote" system for shared reflection and dialogue, but some pupils voiced concerns about privacy or worried their own responses could be deleted (p. 32). In Barrus (2013), increasing students' awareness of their own thinking deficits seemed to cause a temporary drop in motivation (p. 60). In Fößl et al. (2016) and Kramarski and Friedman (2014) students experienced frustration with the interface for the computerised tutoring programmes developed by researchers and provided specific feedback for improvement. These examples show the range of cognitive, metacognitive, and affective perspectives students offer on an intervention, which could be useful for optimising it in future research or practice, but only 19 out of 62 reports included such views, and of these more than half reported only positive comments.

Regarding teachers' views of the interventions, 46 out of 62 reports did not include any responses, six included only positive responses, another six included both positive and negative responses, and four reported only negative responses. While some stated in a general way that teachers found the interventions acceptable or useful, others offered more focused feedback. In Collingwood and Dewey (2018), a teaching assistant who delivered the intervention expressed that the trained affective-regulation and task-approach strategies would also be beneficial in contexts beyond mathematics learning (p. 86). This touches on the idea of "transfer," which has been frequently discussed in MC/SRL research, and it connects to an early assertion by Flavell (1979) that metacognition would lead to better life choices in addition to better school achievement (p. 910). Some teachers also pointed to benefits for students like improved behavioural regulation (Ford, 2018), motivation (Fößl et al., 2016), and resourcefulness (Byrd, 2019). Interestingly, in Cross (2009, p. 926), teachers noted that the written and verbal argumentation exercises gave them insight about students' current misconceptions. Thus, the MC/SRL activities could help strengthen the feedback loop of teaching and learning within the class. Not all intervention leaders' responses were positive, however. Teachers sometimes experienced difficulty implementing the MC/SRL training as intended or on schedule. Barrus (2013, p. 60) states that classroom management became more difficult as pupils were expected to work with learning software independently, rather than having the teacher directly lead the lesson. In Byrd (2019, pp. 88-89) there were frequent technology-related issues that impeded the intervention progress. Edwards (2008) reports that teachers in the intervention group thought the pace of activities was too rushed (p. 77) and that one was demoralised about taking time away from test preparation and did not continue using the intervention “with the same resolve” (p. 102).

While such concerns could arise with a variety of interventions, there could be some unique to MC/SRL interventions in mathematics. For example, Cross (2009, p. 926) found teachers to struggle with adopting new classroom norms based on their experiences of “traditional” teaching and beliefs about the nature of mathematics learning. This is noteworthy because MC/SRL programmes encourage students to take a more self-directed approach to learning, while giving them the tools and opportunities to do so. Compared with a traditional, teacher-centred approach, MC/SRL programmes could require major shifts in classroom culture to be implemented successfully, as Cross points out. In Sings Jenkins (2009, p. 44), one intervention teacher was removed from the study because she did not fully implement some of the SRL strategies she found challenging, indicating a need for better training or support in the future. Teachers’ beliefs about their students were also illustrated in some responses, as shown in Ford (2018). In that study, the teacher expressed doubt beforehand that students would be able to meet the expectations of the new programme in their behaviour or task performance, but these concerns were not realised (p. 73). In other reports, teachers also voiced concern that students with language-related or other special needs might not benefit from the interventions to the same extent as others (Motteram et al., 2016, p. 29), or that the mathematical tasks used in the intervention could be too difficult (Vula, 2017, p. 57). While this reveals how MC/SRL training might be accepted or rejected by teachers, no trends can be identified due to missing information across the review set. Of the 46 reports that did not include teachers' views, around half reported studies at least partly led by the regular classroom teachers. Even for those interventions led by other individuals, it could be beneficial to note what the classroom teachers knew about the interventions and how they responded, but this has been largely overlooked[32]. It would be expected that the success of MC/SRL strategies relies, to some extent, on students’ and teachers’ willingness to use them. Using such strategies perfunctorily without being convinced of their value would undermine a MC/SRL rationale based on explicit and strategic thinking. Ignoring participants’ views could also add weight to criticisms of following the “medical

[32] In some cases, measures were taken to obscure the goals of the research from teachers (e.g., Desoete, 2009, p. 442), though strict blinding/masking protocols were not often reported.

model" (Morrison, 2021, pp. 103-105) of effectiveness in education research, rather than considering teaching and learning from a more holistic, community-based perspective.

Response	**Student responses**	**Teacher responses**	**Student and teacher responses**
Positive only	10	6	9
Negative only	1	4	4
Mixed	8	6	12
Unclear	43	46	37
Total reports	62	62	62

Table 18. Codes applied to included reports for social acceptability of the MC/SRL intervention.

4.1.10 Mathematics content and the MC/SRL interventions

I next examined the mathematical content of the teaching for included studies. If not explicitly given, this could sometimes be determined from the assessment descriptions. Only some reports included samples of the mathematical tasks students worked with, and a few reports included a full list of the mathematics topics. Not all interventions were focused on mathematics specifically since all studies with a mathematics skills outcome were eligible for inclusion. Instead, interventions could focus on general MC/SRL strategies for learning and include mathematics tasks only occasionally or not at all. The mathematical operations involved were not always clear. Thus, the codes represent different levels of specificity and are not mutually exclusive. For example, Collingwood and Dewey (2018) and Jackson Jackson (2012) both included tasks related to money and time. In Bruce (2015) students were trained in goal-setting and completed individualised plans to prepare for the Measures of Academic Progress (MAP) assessment, therefore different students would likely be exposed to different mathematical content, but this is not described in the report. Figure 15 shows the mathematical content codes applied. Although studies could involve multiple types of tasks or mathematics content, many had a specific focus, such as on word problems (19 studies). After word problems, the most common mathematics areas used were multiplying/dividing (15 studies), adding/subtracting (13 studies), algebraic functions (13 studies), and geometry or trigonometry tasks, including those examining properties of circles (11 studies). The remaining codes were all used in fewer than 10 studies. Some studies mentioned choosing a specific task or topic to intervene upon due to its difficulty for many pupils to master. For example, word problems were said to represent a challenge due to their use of specialised language and the need to infer the necessary operations rather than have them clearly determined by the problem statement. Morales (2016) worked with English learners who needed assistance to understand such problems and explain their approaches. Wang et al. (2019) focused on fractions, which they argue has not received enough focus as a “foundational” (p. 339) skill with which many students struggle. Although the students in Wang et al. (2019) had documented “difficulties” in mathematics, many included studies did not make clear whether their participants found the target tasks hard or easy. Difficulty could be relevant to the expected usefulness of MC/SRL approaches. When students experience difficulty (e.g., high effort, low achievement, negative social comparisons), this would be viewed as important feedback on self-efficacy within the social-cognitive model of self-regulation (Bandura, 1991). Depending on how learners interpret this feedback, they may be encouraged to expend more effort or strategic behaviour or may be discouraged from doing so. As pointed out in the EEF’s guidance report on metacognition and self-regulated learning (Quigley et al., 2018, p. 18), goals or tasks should present a rewarding but reachable challenge that is just beyond one’s current capabilities, and MC/SRL strategies could be most effective in a context of optimised difficulty. Yet to evaluate the empirical evidence for this model, and to confirm the role of MC/SRL strategies, would require more primary-study reporting about students’ perceptions of difficulty.

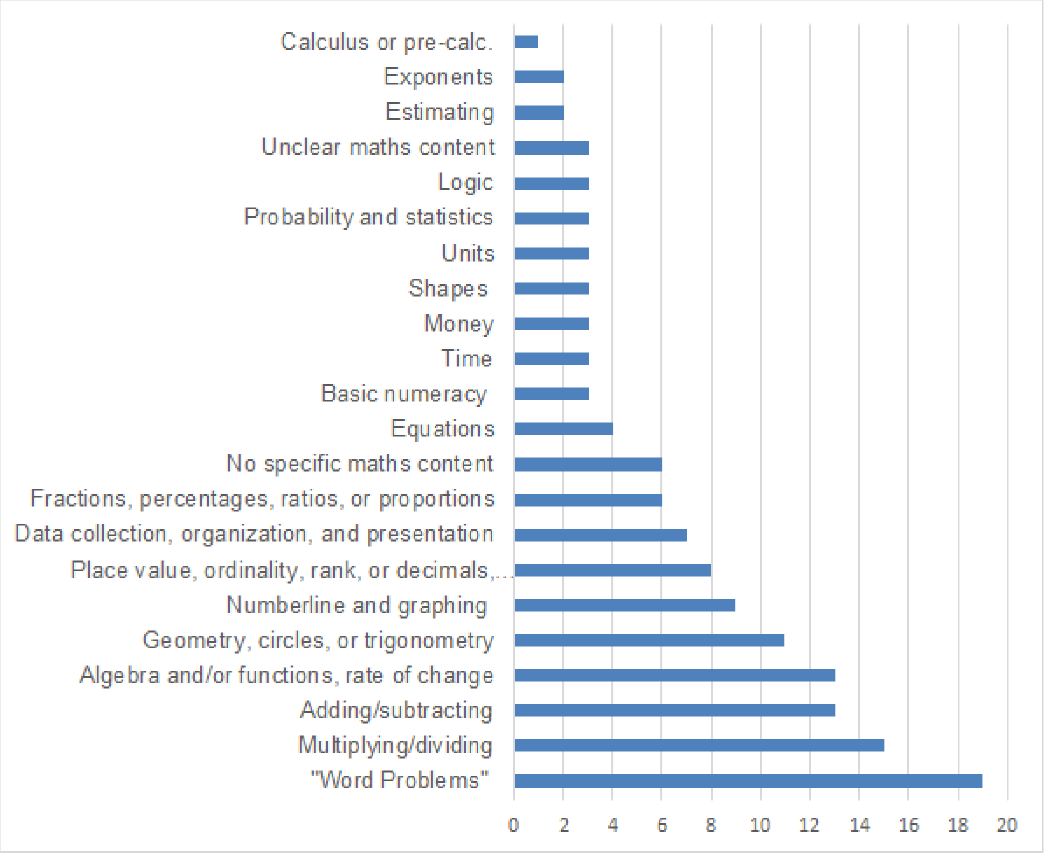


Figure 15. Codes applied to studies for mathematical content. Studies could receive multiple codes.

In some cases, the interventions were designed to accompany either a standard unit from the curriculum or a newly developed unit also being used in the comparison group, but without MC/SRL elements (e.g., Bond & Ellis, 2013; Edwards, 2008). In other cases, teachers were instructed to use the MC/SRL strategies in normal lessons (e.g., Baliram & Ellis, 2019). When the intervention involved solving problems on a computer or other device, the content of these problems could be tailored to the curriculum of the class or broad-ranging, and the SRL elements could be embedded in the problems themselves or used alongside them. To determine how the MC/SRL training aligned with the mathematics content, studies were coded for their mathematics "embeddedness" (see Table 19). From 60 studies, 51 were coded as having MC/SRL training taught through mathematics–that is, it was presented as being woven through mathematics learning, such as planning for, monitoring, and reflecting on mathematical tasks or learning about general mathematics principles. The IMPROVE studies, and others in which a structured approach to problem-solving were taught, fit this category (e.g., Babakhani, 2011; Cornoldi et al., 2015; Jitendra et al., 2015; Tok, 2013). Aside from self-questioning, studies with this code taught strategies for defining and representing tasks, such as through schematic drawings (e.g., Abdullah, 2014; Hughes et al., 2019), explaining task approaches to others verbally (e.g., Finau et al., 2018) or in writing (e.g., Morales, 2016), or controlling attention and affect while completing tasks (e.g., Perels, Dignath, & Schmitz, 2009). The intervention needed to be designed to facilitate learning and performance in mathematics specifically to receive this code, and the data extraction shows that most of the included studies had such a focus. Twelve studies trained MC/SRL skills alongside mathematics. The interventions focused on, for example, behavioural and emotional regulation, performance-related goal setting, or general reflection on learning (Motteram et al., 2016). As studies were done in schools, students were involved in normal mathematics classes, but there was no clear connection between the MC/SRL activities and the mathematics learning. In several cases, there were multiple elements to the intervention, where some included mathematical tasks and some did not, and then both codes were used. Rarely, the MC/SRL training was done through a non-mathematical domain or outside of an academic class (six studies). Mathematics skill-assessment needed to be a pre-planned outcome for reports to be included in the review. Still, the MC/SRL training was sometimes delivered in a language arts class (e.g., Hughes et al., 2019) or during a play period (e.g., Tominey & McClelland, 2011). If the intervention included activities in both academic and non-academic contexts, this also resulted in multiple codes being applied.

Level of mathematics embeddedness	**Number of codes**
MC/SRL taught through mathematics	51
MC/SRL taught alongside mathematics	12
MC/SRL taught through other or no academic subject	6

Table 19. Codes applied to studies regarding mathematical embeddedness of the interventions. Note that some studies received two codes.

4.1.11 Theoretical discussions and rationales for MC/SRL programmes

All studies in the current review needed an explicit basis in MC/SRL theory to be included. In addition, most studies mentioned other related theories and concepts, such as motivation, self-efficacy, and self-reflection. Several previous reviews considered primary studies' theoretical basis and, in some cases, used this information to categorise the interventions (e.g., Dignath, Büttner, & Langfeldt, 2008; Wang & Sperling, 2020). For the current review, I examined the theoretical discussions of included reports. I was also interested to see whether student and teacher participants in the studies were provided with a rationale for using MC/SRL strategies including a discussion of MC/SRL theories. This second point has not been covered in previous reviews, but I considered it relevant, since MC/SRL programs may work directly through participants' beliefs, motivations, and informal theories of mind. During the data extraction, I found that most reports included a detailed rationale for the MC/SRL intervention based on empirical and theoretical research and information about the target context. Three-fourths of studies (n=44) reported rationale in "high detail," while the remaining featured "moderate detail." However, discussions of theory and rationale were less clear for participants, with only a third of studies (n=21) reporting they discussed theory or rationale with students, while half (n=29) discussed this with teachers or other intervention leaders. This latter number includes studies where researchers led the intervention and knowledge of the underlying theories is assumed where not explicitly stated.

Was the theoretical basis or rationale discussed . . .			
in the report?	high detail	moderate detail	low detail
	44	16	0
with pupils?	yes	no	unclear
	21	0	39
with teachers or other intervention leaders?	yes	no	unclear
	29	0	31

Table 20. Codes applied to included studies for discussion of theoretical basis or rationale.

There was great variety in the types of MC/SRL theories or sub-theories alluded to in the study rationales, with many reports mentioning multiple theories and researchers, and no clear patterns across the review. Similar to other reviews, I found that general metacognitive, social-cognitive, and motivational theories were discussed in multiple studies, in addition to more limited concepts such as "maths anxiety" (Collingwood & Dewey, 2018), attributions (Dresel & Haugwitz, 2008), calibration (Riggs, 2012), executive function (Schmitt, 2013), and opportunity-to-learn (Wijaya et al., 2018). Concepts specific to mathematics were also discussed, such as the use of consistent or inconsistent language in problem statements (Mevarech et al., 2010), Pólya's phases of problem-solving (Morales, 2016; Lee, 2014), and cognitive conflict (Finau et al., 2018). Where interventions focused on understanding and representing mathematical tasks (Pennequin et al., 2010), choosing effective strategies (Shilo & Kramarski, 2019), and reflecting on errors (Heemsoth & Heinze, 2016), rationales more explicitly linked MC/SRL theories with the specific needs of mathematics learners. In some other cases, MC/SRL activities were presented as generally beneficial for learning in school without clearly locating this within the mathematics domain (e.g., Motteram et al., 2016). For example, Bond and Ellis (2013) trained students in “reflective assessment” and a “think aloud” technique designed to encourage metacognitive growth, yet there is no discussion of how this would be beneficial in mathematics specifically. Treatment of theories within MC/SRL studies is discussed further in the final chapter.

4.1.12 The MC/SRL stages, activities, and strategies

Because specific mathematical tasks often set the context for the MC/SRL training, I report the “stage of MC/SRL” on which the intervention focused. Boekaerts and Corno (2005) refer to “orientation, performance, and verification stages of mathematics problem solving” (p. 2011). The idea is that meta-level skills take different forms when they are used before, during, or after a task. For the current review, many reports alluded to models of SRL or problem-solving that included three or four stages, one or two prior to task performance, one during it, and one after it. Rather than categorising studies based on such models, I recorded whether they focused on pre-task, during-task, or post-task strategies, or a combination. It was found that 47 out of 60 studies included multi-stage or general MC/SRL strategies, while four were considered only pre-task interventions, four during-task, and five post-task. Pre-task interventions included activities such as setting performance or learning goals, understanding tasks, and planning a task approach. Calibration, or predictions about task performance (e.g., Riggs, 2012), were also considered pre-task activities. During-task interventions focused on monitoring and control of cognition, attention, behaviour, or affect, and generally used specific prompts or questions to guide learners. For example, Shamir and Lifshitz (2013) used e-books that included prompts on each page to help students monitor their thinking. Post-task activities included reflections on learning or performance and making appropriate attributions for successes and failures. Heemsoth and Heinze (2016) had pupils consider the faulty reasoning or mistakes in calculation behind their own errors. As noted, more than three-fourths of the interventions included strategies for multiple task stages, or they utilised general approaches, such as verbalising thought processes, that could be employed at any stage of a task.

In addition to the stage of MC/SRL, the current review analysed interventions in a more fine-grained way by extracting information about the specific activities and strategies each contained, shown in the "Activities and Strategies" table in Appendix 10. As above, these were coded based on what was done and said in the interventions, not on the theoretical discussion in the reports. Because a single intervention activity could have multiple components, different combinations of activities and strategies were seen. Some of the codes related more to the mode of learning, such as writing, discussion, or graphing, while others related more to the MC/SRL focus, such as planning, monitoring, or reflecting on learning.

Activity or Strategy Used	**Studies**
Questioning, monitoring, or control of learning or task processes or performance	48
Mathematics problem-solving	45
Defining or planning for tasks or learning activities, strategy choice	44
Self-evaluation or self-prediction (e.g., calibration) of knowledge/performance	42
Discussion (verbal or written)	39
Explaining task approaches/TA (thinking aloud)/ students teaching peers, recording audio reflections	35
Providing or receiving/reviewing marks or feedback, error-correction	34
Graphing, modelling with images, taking photographs, or colouring	28
Writing about thinking	26
Learning/reviewing mathematics principles	22
Affective/motivational regulation	17
Behavioural, attentional, time, and environment regulation	16
Setting learning or achievement goals	15
Use of play, games, music/sounds, drama, or humour	11
Memory and "study" strategies, note-taking	9
Mathematical language/terms and reading strategies	9
Unstructured mathematics exploration, task formulation or choice by students	8
Construction or manipulation of physical props	6
Attributions for performance	5

Activity or Strategy Used	Studies
Physical exercise, movement, or breathing	4

Table 21. Activities and strategies used in interventions and number of reports with each code.

As shown, the most common intervention activity was “questioning, monitoring, or control of learning or tasks” (48 studies). This is not surprising given the large number of studies using the IMPROVE approach or a similar self-questioning scheme. Just as IMPROVE uses a mnemonic to cue essential actions when problem-solving, other interventions also used a memorable word or phrase or pictorial guide to aid students in remembering the steps (e.g., Lee, Yeo, & Hong, 2014). This code was also applied when students received external prompts during problem-solving, such as from the teacher or a computerised tutoring programme. “Mathematics problem-solving” was the second most common code here, being used by 45 studies out of 60. This was only coded when, as part of the intervention, students worked on tasks requiring mathematical computations, such as answering word problems or solving for a variable in an equation. Tasks in which students were not expected to produce a correct or optimal solution did not receive this code. The next most common code (44 studies) was “defining or planning for tasks.” Interventions with this code trained students in strategies for reading and understanding word problems such as by identifying relevant versus irrelevant information, modelling tasks with graphs or other images, identifying the requisite operations to reach a solution, and choosing among different potential solution strategies. Another common code (42 studies) was used when students evaluated their own knowledge and skills relative to a task, either before or after attempting it. Calibration exercises, which asked pupils to predict their score on an assessment tool, were also included in this category (e.g., Riggs, 2012). These were also combined with post-task self-evaluations, either before or after receiving correctness feedback. More than half of included studies (39) stated they used some kind of discussion around learning to encourage MC/SRL processes. Pupils explained their own reasoning and prompted, questioned, or challenged that of their peers, but sometimes the discussion was less specified in reports. The “discussion” code was only used for short, back-and-forth exchanges in person or mediated by technology, whereas there was a different code for written explanations of thinking that did not presume a reply (26 studies). Thirty-four studies included activities relating to feedback on performance or correctness, whether on a test, or from a teacher or peer. Studies that focused on strategies for error-correction also received this code, since students had to be notified about their errors to correct them. Sometimes the correctness feedback was delivered by a computer-based tutoring programme, along with prompts for re-thinking the task or solution (e.g., Kramarski & Gutman, 2006).

There was also a code for activities and strategies that utilised verbal or written explanations of one's own thinking (35 studies). Interventions using the term “think-aloud” were included in this category. Motteram et al. (2016) had pupils record their reflections in video or audio. Any activity where students explained or justified their task approaches fit in here. Another code (28 studies) was used for activities in which students generated images representing a task or their own learning, such as through photographs or drawings. Frequently, these images displayed relationships between elements in a problem (e.g., Abdullah, Halim, & Zakaria, 2014; Mandaci Şahin & Kendir, 2013), which could free students from only focusing on the numeric operations required. These images could also demonstrate whether students were properly interpreting information in a word-problem, for example, so that they did not start on the wrong approach.

The remaining activities and strategies were used in less than half of the included studies. They included (from most common to least common): learning or reviewing mathematics principles; regulation of affect or motivation; regulation of behaviour, attention, time or the learning environment; setting learning or achievement goals; use of play, games, music, drama, or humour; memory, note-taking, and study strategies; mathematical language and reading strategies; unstructured mathematics exploration or self-formulation of tasks; construction or manipulation of physical props; performance related attributions (self- or other-focused); and physical movement or breathing exercises. In Appendix 10, examples for each of the activities and strategy codes are given from the included studies.

The materials used in the interventions are shown in Table 22, although this information was not always included in the reports. Because interventions could have multiple components with different materials, several codes could be used for the same study. Texts, either physical or digital, were used in most studies (n=46), while nearly half (n=28) reported using images of some kind. Personal digital devices were used in only 12 out of 60 studies, while in eight studies the intervention involved passive video or audio display, and in four studies “smartboards” or “multi-touch” devices were used. Items for creating crafts (six studies) or drawing tools other than pencils (two studies), were used rarely. Cards, coins, or other manipulatives and consumable items were used in three and two studies, respectively. In six studies, the materials used were not reported clearly.

Intervention Materials	**Studies**
Books, papers, notecards, or screens with text	46
Pictures, images, diagrams, posters	28
PCs, laptops, tablets, phones (personal devices)	12
Passive video or audio display, slide show	8
Paint, coloured or craft paper, stickers, glue, scissors	6
Whiteboard, smartboard, or "multi-touch" device	4
Board games, playing cards, dice, coins, or manipulatives	3
Rulers, protractors/compasses, or other measurement/drawing tools	2
Gum, candy, or other consumables	2

Table 22. Intervention materials coded and number of studies receiving each code.

Some reports (e.g., Kang, 2010) emphasised that MC/SRL training could be accomplished with simple materials, such as worksheets. I had planned to extract information about the interventions’ costs, but these were not mentioned in most reports. Only Motteram et al. (2016) reported a detailed cost estimate of the ReflectED at £18.72 per pupil per year, for a three-year implementation. While this estimate assumed schools already had iPads or other devices with which to take photographs and record reflections on learning, the estimate included the cost of paper and other physical materials, training for intervention leaders, and a subscription to the Evernote programme (p. 5). The authors note that the Education Endowment Foundation (EEF), which sponsored the study, rates this as a very low-cost educational intervention. Although the cost of training intervention leaders or purchasing software licences could add to the overall cost of an intervention, it is likely that the cost of materials would be low for most of the interventions in this review. This review demonstrates that there are a range of approaches to training MC/SRL skills, most of which would be extremely affordable.

4.1.13 Quality of evidence

While Dent and Koenka (2016) included intervention and observation studies in their review, I sought to include only designs that could indicate causal relationships between the MC/SRL programmes and mathematics outcomes and thus contribute to an understanding of effectiveness. This cannot be demonstrated, however, unless potential confounds and biases are limited (i.e., “threats to validity,” Hedges, 2012, p. 28). To gauge this, I judged included studies based on specific questions regarding their quality control factors, as shown in Table 23. Included reports varied widely in how participants were selected and assigned to groups, whether they were informed about the study and their group assignment, and the nature of the comparisons and assessments conducted. I found that two-thirds or more of reports did not make it clear: whether participants were blinded or masked (42 studies); whether there was potential selection bias (39 studies); whether participants could have switched treatments during the study (46 studies); whether participants' beliefs, such as resentful demoralisation, could have impacted the results (47 studies); and whether the MC/SRL programmes were implemented with high fidelity (44 studies). It was also frequently unclear whether groups were equivalent at baseline (29 studies) or had low or equivalent attrition levels (20 studies). Most studies did not evaluate outcomes appropriately, given the allocation to study conditions (39 studies). That is, many studies allocated schools, classes, or small groups to study conditions, but they evaluated and reported student-level outcomes without any adjustment for nested or non-independent data. This was also unclear when there was no information about how students were allocated to groups in the first place. Almost one-third of studies did not use standard teaching as the comparison for the MC/SRL-based programme (19 studies), and a quarter of studies used more than one mathematics assessment (15 studies) or an assessment that was not developed separately and validated prior to the study (14 studies). Use of multiple assessments or analyses could result in spurious findings (Gorard, 2014, p. 53), and use of unvalidated assessments or those validated with the same data used to report the study outcomes could undermine reliability. In total, 11 quality-control factors were examined, and on average these were implemented and reported appropriately by only one-third of primary studies (21 studies).

Question or item	**Yes**	**No**	**Unclear**
Blinding or masking performed?	13	5	42
Low or no selection bias?	15	6	39
Baseline equivalence demonstrated?	24	7	29
Business-as-usual comparison used?	35	19	6
Evaluated at level of allocation to condition?	14	39	7
Only one mathematics assessment used?	44	15	1
Mathematics assessment separately developed/validated?	29	14	17
Low or balanced attrition/missing data?	28	12	20
Low contamination or crossover?	12	2	46
No evidence of participants' beliefs affecting outcomes?	2	10	47
High fidelity of treatment?	13	3	44
Total	229	132	298
Average	21	12	27

Table 23. Codes applied to studies regarding quality control and strength of evidence.

4.1.14 Ethical issues

Ethical participation factors were also examined for the narrative synthesis. Although ethics standards for research vary, many identify the need for informed consent or voluntary participation. Information about how personal data will be used should also be communicated to participants. Since educational research frequently involves vulnerable children and adolescents, researchers must be sensitive to even minor harms, such as emotional stress or missed learning opportunities. On the other hand, explicit consent may not be required in all localities, especially when the interventions and assessments are similar to those already in use and only aggregated data is reported. Teachers could also suffer participation harms, such as pressure to participate in a trial or reprisals for poor performance with the intervention. I was interested in how studies for this review negotiated ethical issues, but reporting was generally lacking. Most studies did not report how students or teachers were enrolled in the study or how personal data was stored or protected (see Tables 24 and 25). In fact, in several cases, samples of student work were used in the primary study reports without de-identification (Edwards, 2008; Tok, 2013). Some studies, while not identifying teachers, reported their engagement in the study in a negative light that could lead to consequences if they were identified by superiors. For example, Cross (2009) reports:

> . . . teachers took a fairly traditional, teacher-centered approach to teaching and so in order to facilitate [discussion] they had to . . . relinquish some of the control for knowledge building to the students as well as manage a classroom that was active and alive with students' talk and movement. Having to function in this way seemed to threaten the perceptions they had of their teacher role. . . .and how they thought students learned best. (p. 926)

As illustrated in this quote, MC/SRL interventions may necessitate novel classroom dynamics and upset teachers' and pupils' beliefs about learning. While voluntary participation might indicate a readiness to make such shifts, Cross (2009) does not state whether teachers had a choice to participate or not, and this could be important for understanding their negative reactions and lack of intention to continue using the intervention (p. 927). Based on the lack of reporting in the included studies, no trends in ethical participation can be outlined for the present review.

Question	**Yes**	**No**	**Unclear**
Consent/assent for the study sought from students themselves?	15	0	45
Consent sought from parents/guardians?	16	1	43
Personal data protected?	12	2	46
Students free from negative consequences/reprisals?	6	0	54

Table 24. Codes applied to studies regarding ethical participation of students.

Question	**Yes**	**No**	**Unclear**
Consent sought from teachers/staff?	13	0	47
Personal data protected?	5	2	53
Teachers/staff free from negative consequences/reprisals?	2	1	57

Table 25. Codes applied to studies regarding ethical participation of teachers and staff.

4.2 The meta-analysis results

This section presents results from the posttest-only and pre/post meta-analysis of study-level effects on mathematics skills or achievement outcomes. As a reminder, each study contributed a single effect size to each meta-analysis based on grouping together all MC/SRL study conditions and comparing them to all non-MC/SRL conditions, according to the methods described in the previous chapter. Study-level effects were generated using means, standard deviations, and sample sizes, where reported, with any necessary divergence from this detailed above and in Appendix 7. Effect sizes and standard errors for all 60 included studies were entered into a Meta-Essentials (Suurmond, van Rhee, & Hak, 2017) workbook, yielding results for the meta-analyses, heterogeneity check, publication bias check, and moderator and subgroup analyses. These results are presented and interpreted here, with further discussion in the next chapter. Consideration is given to how these results can be applied to teaching and learning practices and future research.

4.2.1 The posttest-only meta-analysis

The forest plot in Figure 16 shows the results of the posttest-only meta-analysis including outcomes from 60 primary studies. Recall that some studies only reported adjusted outcomes, so in a few cases these effects control for baseline differences. Studies are ordered by weight descending, so that those at the top have the most impact on the overall combined effect size. The individual study effects are shown as blue circles with the size of the circle indicating the weight based on the variance. Each individual effect also appears with its 95% confidence interval shown as black bars. As shown in the Figure 16, the studies with the lowest variance have the narrowest confidence intervals and the highest weight in the combined effect size calculation. Although the two most highly weighted studies, Jitendra et al. (2015) and Motteram et al. (2016) both have negligible effects, the overall combined effect size is 0.46, with a standard error of 0.08. This is shown by the green circle at the bottom of the plot and is also reported in Table 26. The 95% confidence interval for the combined effect size, represented by the black bars next to the green circle, ranges from 0.30 to 0.63. Using the traditional language of statistical significance, this means the combined effect is “significant,” or not considered likely to have occurred by chance, since the confidence interval excludes the null, though the problems with this interpretation have been discussed at length (e.g., Gorard, 2015). The green bars extending on either side of the black bars represent the prediction interval, which is sometimes interpreted as indicating the range of potential future studies (IntHout, et al., 2016).

Due to lack of consensus on its meaning, only the prediction interval for the posttest-only meta-analysis is reported.

Study	ES	95% CI LL	95% CI UL	Weight
Jitendra et al (2015)	0.05	-0.04	0.14	2.18%
Motteram et al. (2016)	-0.01	-0.11	0.09	2.17%
Shilo & Kramarski (2019)	0.39	0.26	0.53	2.14%
Wijaya et al. (2018)	0.21	-0.02	0.44	2.04%
Finau et al. (2018)	1.03	0.79	1.27	2.03%
Vula et al. (2017)	0.18	-0.07	0.42	2.02%
Schmitt (2013)	0.08	-0.18	0.33	2.01%
Bruce (2015)	0.43	0.16	0.69	2.00%
Ubuz & Erdoğan (2019)	-0.08	-0.35	0.19	1.99%
Abdolhossini (2012)	0.60	0.32	0.89	1.96%
Mevarech et al. (2010)	0.56	0.26	0.86	1.95%
Heemsoth & Heinze (2016)	0.22	-0.08	0.52	1.94%
Abdullah, Halim, & Zakaria (2014)	1.11	0.80	1.41	1.93%
Cross (2009)	0.36	0.05	0.67	1.92%
Falco (2008)	0.00	-0.32	0.32	1.91%
Collingwood & Dewey (2018	0.37	0.04	0.70	1.89%
Tzohar-Rozen & Kramarski (2017)	0.80	0.46	1.14	1.88%
Cornoldi et al. (2015)	0.10	-0.24	0.45	1.87%
Kramarski, Weisse, & Kololshi-Minsker (2010)	0.72	0.38	1.07	1.87%
Jackson Jackson (2012)	0.18	-0.17	0.53	1.86%
Dresel & Haugwitz (2008)	0.79	0.44	1.15	1.85%
Lestari & Jailani (2018)	0.43	0.07	0.79	1.84%
Bond & Ellis (2013)	0.92	0.55	1.29	1.83%
McClelland et al. (2019)	-0.04	-0.41	0.33	1.82%
Riggs (2012)	-0.02	-0.40	0.36	1.81%
Kramarski & Dudai (2009)	0.44	0.02	0.86	1.75%
Arroyo et al. (2007)	0.08	-0.38	0.54	1.68%
Kramarski & Friedman (2014)	0.91	0.45	1.38	1.67%
Kramarski & Zoldan (2008)	1.22	0.75	1.68	1.66%
Chen & Chiu (2016)	0.07	-0.40	0.54	1.66%
Baliram & Ellis (2019)	0.56	0.09	1.03	1.65%
Aminah, et al. (2018)	0.25	-0.22	0.73	1.64%
Shamir & Lifshitz (2013)	0.71	0.22	1.20	1.62%
Fößl et al. (2016)	0.84	0.34	1.33	1.61%
Kramarski & Gutman (2006)	0.44	-0.07	0.94	1.60%
Tominey & McClelland (2011)	0.32	-0.19	0.82	1.60%
Lee, Yeo, & Hong (2014)	0.25	-0.26	0.75	1.60%
Mevarech & Amrany (2008)	0.38	-0.13	0.90	1.58%
Morales (2016)	1.00	0.48	1.51	1.58%
Desoete (2009)	1.14	0.61	1.67	1.56%
Kang (2010)	0.87	0.34	1.40	1.55%
Mandaci Şahin & Kendir (2013)	1.61	1.08	2.14	1.55%
Babakhani (2011)	0.49	-0.05	1.02	1.55%
Edwards (2008)	-0.40	-0.96	0.15	1.52%
Sings Jenkins (2009)	0.62	0.07	1.18	1.51%
Perels, Dignath, & Schmitz (2009)	0.44	-0.12	1.00	1.51%
Jacobse & Harskamp (2009)	0.50	-0.09	1.10	1.45%
Pappas Schattman (2005)	0.49	-0.13	1.11	1.41%
Wang et al. (2019)	1.00	0.37	1.63	1.39%
Tok (2013)	1.86	1.21	2.51	1.36%
Pennequin et al. (2010)	1.50	0.82	2.17	1.33%
Cleary, Velardi, & Schnaidman (2017)	-0.35	-1.03	0.33	1.33%
Sarette (2014)	0.13	-0.57	0.83	1.30%
O'Neal (2015)	-1.21	-1.92	-0.50	1.28%
Barrus (2013)	0.54	-0.17	1.25	1.27%
Ozsoy & Ataman (2009)	2.00	1.28	2.72	1.25%
Ford (2018)	-1.69	-2.52	-0.86	1.11%
Hughes et al. (2019)	0.23	-0.61	1.07	1.10%
Byrd (2019)	-1.06	-1.96	-0.16	1.04%
Rizk, Attia, & Al-Jundi (2017)	2.84	1.94	3.75	1.01%

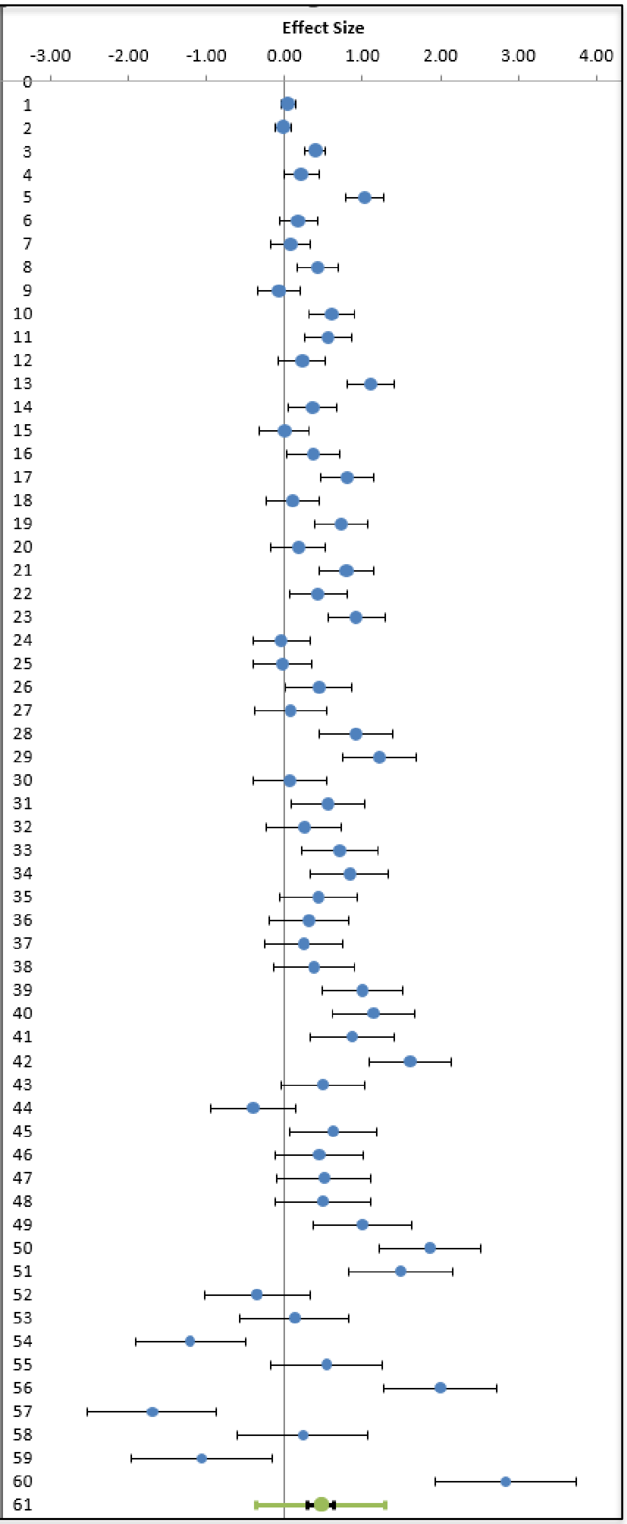


Figure 16. Posttest-only meta-analysis forest plot, studies ordered by weight.

Combined effect size	0.46
Standard error	0.08
95% Confidence interval	0.30 to 0.62
Prediction interval	-0.37 to 1.30

Table 26. Results of the posttest-only meta-analysis.

4.2.2 Publication bias estimation

A chief concern for meta-analysis is whether the combined values include all, or a representative fraction, of those that exist in the world. Otherwise, this undermines the validity of the combined effect. One way to check for this "publication bias" (Song, Hooper, & Loke, 2013) is to examine the range of included effects, which would be predicted to lie along a normal curve, with more moderate effect studies and fewer studies with high or low effects. For the current review, if there were imbalance in the range of effects, I would infer that some studies were missed due to their being unpublished or inaccessible to me. I used several helpful tools in Meta-Essentials to check for potential publication bias, but ultimately this is an issue of judgement and is more complex than simply checking a graph. Figure 17 shows a funnel plot of the effect sizes and standard errors of the included studies. The blue circles represent the individual studies, and the green circle and black bars at the bottom the combined effect size and its confidence interval. The red line at the bottom shows the adjusted combined effect size after imputing any "missing" studies, of which there are none here. Publication bias could include null or negative-result studies being unpublished or reported in a lower-access format, leading to asymmetry in the funnel plot. To correct for this potential bias, Meta-Essentials includes a "trim and fill" feature, based on the approach described by Duval and Tweedie (2000), which removes the more "extreme" studies and then replaces them with imputed studies designed to improve the overall symmetry and adjust the combined effect. In this case, there is no clear asymmetry using the original 60 studies of the meta-analysis, so no "missing" studies are imputed and the combined effect size is unadjusted. The range of effects from the primary studies does exceed that expected, as shown by the two red lines on either side of the "funnel." These lines represent the 95% confidence interval for the combined effect given different standard errors. Despite several studies falling outside of these lines, there is no clear positive or negative skew, so these potential "outliers" are retained for all further analyses.

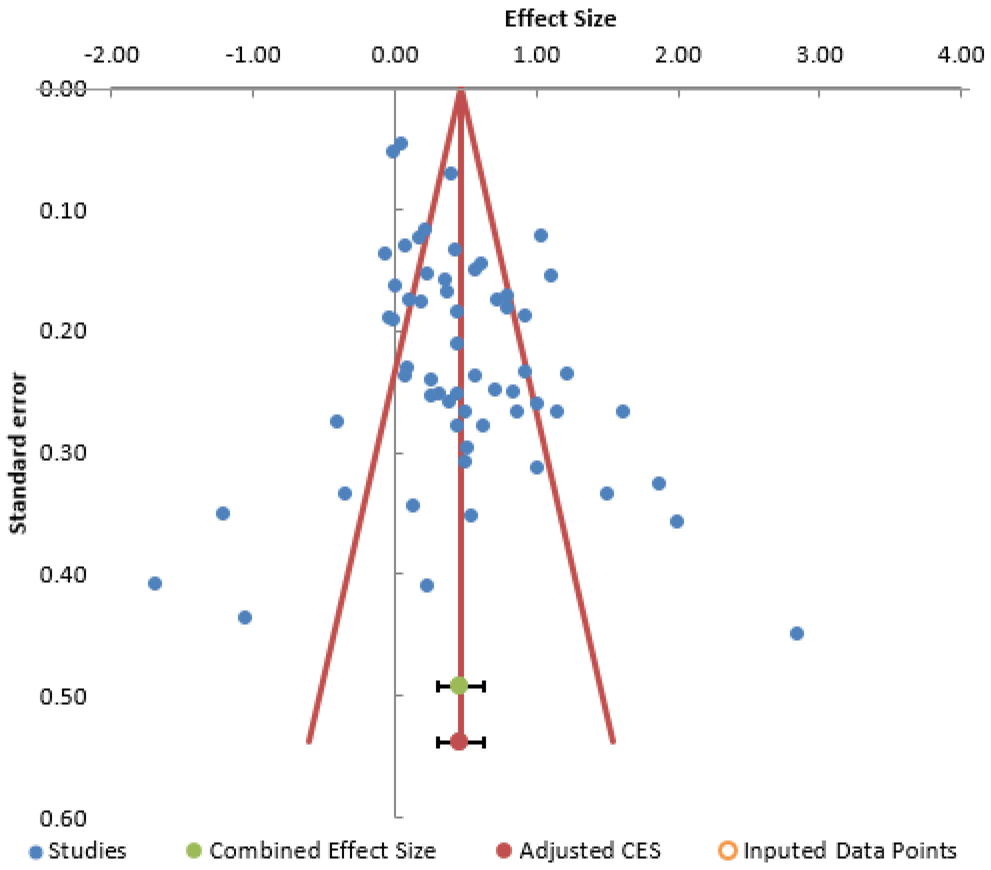


Figure 17. Funnel plot of included study effects and their standard errors.

The interpretation that the observed effects are symmetric around the mean is also supported by an examination of the standardised residual histogram produced in Meta-Essentials, shown in Figure 18. This histogram shows the Z-scores of observed effects are mainly found between 1.5 and 1.5, as expected, without any indication of positive or negative bias.

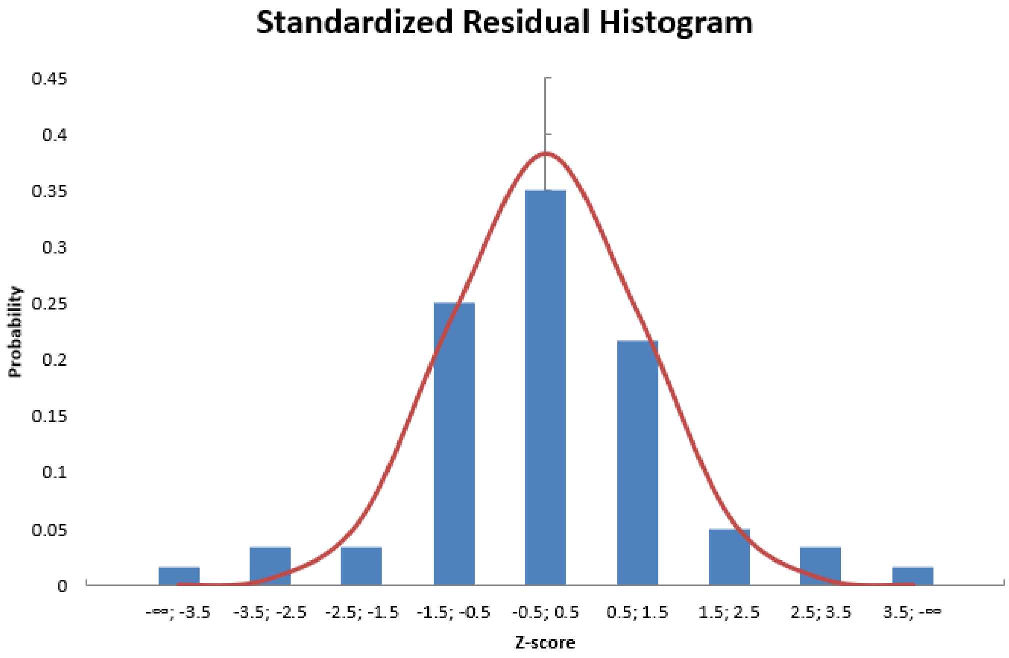


Figure 18. Standardised residual histogram showing expected vs. observed dispersion from the mean of effect sizes.

While the funnel plot and standardised residual histogram examine the symmetry of effects, other analyses within Meta-Essentials estimate the stability of the combined effect based on the included studies. For example, “fail-safe n” tests predict the number of additional studies that would be needed to disturb the combined effect. Several of these are based on the “significance” of the combined effect, whereas the size of the effect is considered of more interest here. Thus, only Orwin’s (1983) test is reported here. Assuming the studies to be added had a mean effect size of 0.0, it is estimated that 498 additional studies would be needed to move the combined effect size to a criterion value of 0.05, essentially a null result. This indicates the combined effect is robust and unlikely to change should more studies fitting the review criteria be identified. Considering all the other analyses above, I judge there to be a low risk of bias in the current review due to unpublished or inaccessible studies.

4.2.3 Heterogeneity in the results

Several values reported in the Meta-Essentials workbooks can be interpreted as markers of heterogeneity (see Table 27). That is, they indicate the extent to which the observed individual study effects do not seem to be sampled from the same population. Although using the random-effects model for meta-analysis assumes that “true” effects of the interventions may vary from study to study, the values reported for this meta-analysis indicate a higher-than-expected variation in effects, indicating it may not have been appropriate to combine the review studies in a single meta-analysis. In fact, the developers of Meta-Essentials, the tool used for this meta-analysis, caution strongly that heterogeneity markers indicate a need for further investigation to identify potential subgroups and moderators of the effect, and that the combined effect size and publication bias analysis will only be meaningful given a “set of homogeneous results” (Hak, van Ree, & Suurmond, 2018, pp. 10, 19). Thus, considering the extent of and reasons for heterogeneity in the meta-analysis are crucial steps.

To begin with, the forest plot of individual study effects in Figure 19 can be visually inspected for signs of heterogeneity (Siebert, 2018). From this, it is evident that the confidence intervals of a large number of the effects overlap, while a number of others do not overlap and indicate heterogeneity. Here, the studies are ordered by descending value of the standard error. Studies shown at the top, those with the largest standard errors, also have less overlap in their confidence intervals. Thus, it could be interpreted that these studies represent the region of greatest heterogeneity.

Study	ES	95% CI LL	95% CI UL	Weight
Rizk, Attia, & Al-Jundi (2017)	2.84	1.94	3.75	1.01%
Byrd (2019)	-1.06	-1.96	-0.16	1.04%
Hughes et al. (2019)	0.23	-0.61	1.07	1.10%
Ford (2018)	-1.69	-2.52	-0.86	1.11%
Ozsoy & Ataman (2009)	2.00	1.28	2.72	1.25%
Barrus (2013)	0.54	-0.17	1.25	1.27%
O'Neal (2015)	-1.21	-1.92	-0.50	1.28%
Sarette (2014)	0.13	-0.57	0.83	1.30%
Cleary, Velardi, & Schnaidman (2017)	-0.35	-1.03	0.33	1.33%
Pennequin et al. (2010)	1.50	0.82	2.17	1.33%
Tok (2013)	1.86	1.21	2.51	1.36%
Wang et al. (2019)	1.00	0.37	1.63	1.39%
Pappas Schattman (2005)	0.49	-0.13	1.11	1.41%
Jacobse & Harskamp (2009)	0.50	-0.09	1.10	1.45%
Perels, Dignath, & Schmitz (2009)	0.44	-0.12	1.00	1.51%
Sings Jenkins (2009)	0.62	0.07	1.18	1.51%
Edwards (2008)	-0.40	-0.96	0.15	1.52%
Babakhani (2011)	0.49	-0.05	1.02	1.55%
Mandaci Şahin & Kendir (2013)	1.61	1.08	2.14	1.55%
Kang (2010)	0.87	0.34	1.40	1.55%
Desoete (2009)	1.14	0.61	1.67	1.56%
Morales (2016)	1.00	0.48	1.51	1.58%
Mevarech & Amrany (2008)	0.38	-0.13	0.90	1.58%
Lee, Yeo, & Hong (2014)	0.25	-0.26	0.75	1.60%
Tominey & McClelland (2011)	0.32	-0.19	0.82	1.60%
Kramarski & Gutman (2006)	0.44	-0.07	0.94	1.60%
Fößl et al. (2016)	0.84	0.34	1.33	1.61%
Shamir & Lifshitz (2013)	0.71	0.22	1.20	1.62%
Aminah, et al. (2018)	0.25	-0.22	0.73	1.64%
Baliram & Ellis (2019)	0.56	0.09	1.03	1.65%
Chen & Chiu (2016)	0.07	-0.40	0.54	1.66%
Kramarski & Zoldan (2008)	1.22	0.75	1.68	1.66%
Kramarski & Friedman (2014)	0.91	0.45	1.38	1.67%
Arroyo et al. (2007)	0.08	-0.38	0.54	1.68%
Kramarski & Dudai (2009)	0.44	0.02	0.86	1.75%
Riggs (2012)	-0.02	-0.40	0.36	1.81%
McClelland et al. (2019)	-0.04	-0.41	0.33	1.82%
Bond & Ellis (2013)	0.92	0.55	1.29	1.83%
Lestari & Jailani (2018)	0.43	0.07	0.79	1.84%
Dresel & Haugwitz (2008)	0.79	0.44	1.15	1.85%
Jackson Jackson (2012)	0.18	-0.17	0.53	1.86%
Kramarski, Weisse, & Kololshi-Minsker (2010)	0.72	0.38	1.07	1.87%
Cornoldi et al. (2015)	0.10	-0.24	0.45	1.87%
Tzohar-Rozen & Kramarski (2017)	0.80	0.46	1.14	1.88%
Collingwood & Dewey (2018	0.37	0.04	0.70	1.89%
Falco (2008)	0.00	-0.32	0.32	1.91%
Cross (2009)	0.36	0.05	0.67	1.92%
Abdullah, Halim, & Zakaria (2014)	1.11	0.80	1.41	1.93%
Heemsoth & Heinze (2016)	0.22	-0.08	0.52	1.94%
Mevarech et al. (2010)	0.56	0.26	0.86	1.95%
Abdolhossini (2012)	0.60	0.32	0.89	1.96%
Ubuz & Erdoğan (2019)	-0.08	-0.35	0.19	1.99%
Bruce (2015)	0.43	0.16	0.69	2.00%
Schmitt (2013)	0.08	-0.18	0.33	2.01%
Vula et al. (2017)	0.18	-0.07	0.42	2.02%
Finau et al. (2018)	1.03	0.79	1.27	2.03%
Wijaya et al. (2018)	0.21	-0.02	0.44	2.04%
Shilo & Kramarski (2019)	0.39	0.26	0.53	2.14%
Motteram et al. (2016)	-0.01	-0.11	0.09	2.17%
Jitendra et al (2015)	0.05	-0.04	0.14	2.18%

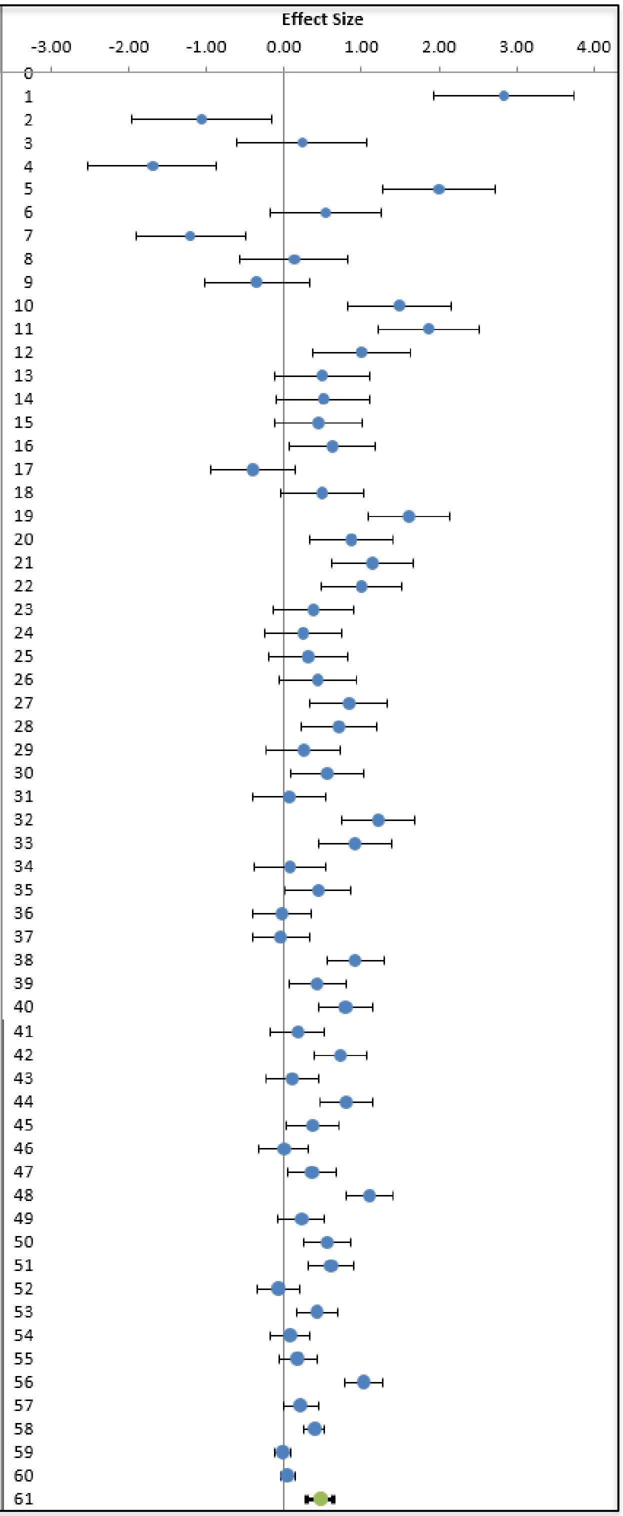


Figure 19. *Posttest-only MA forest plot, studies ordered by standard error.*

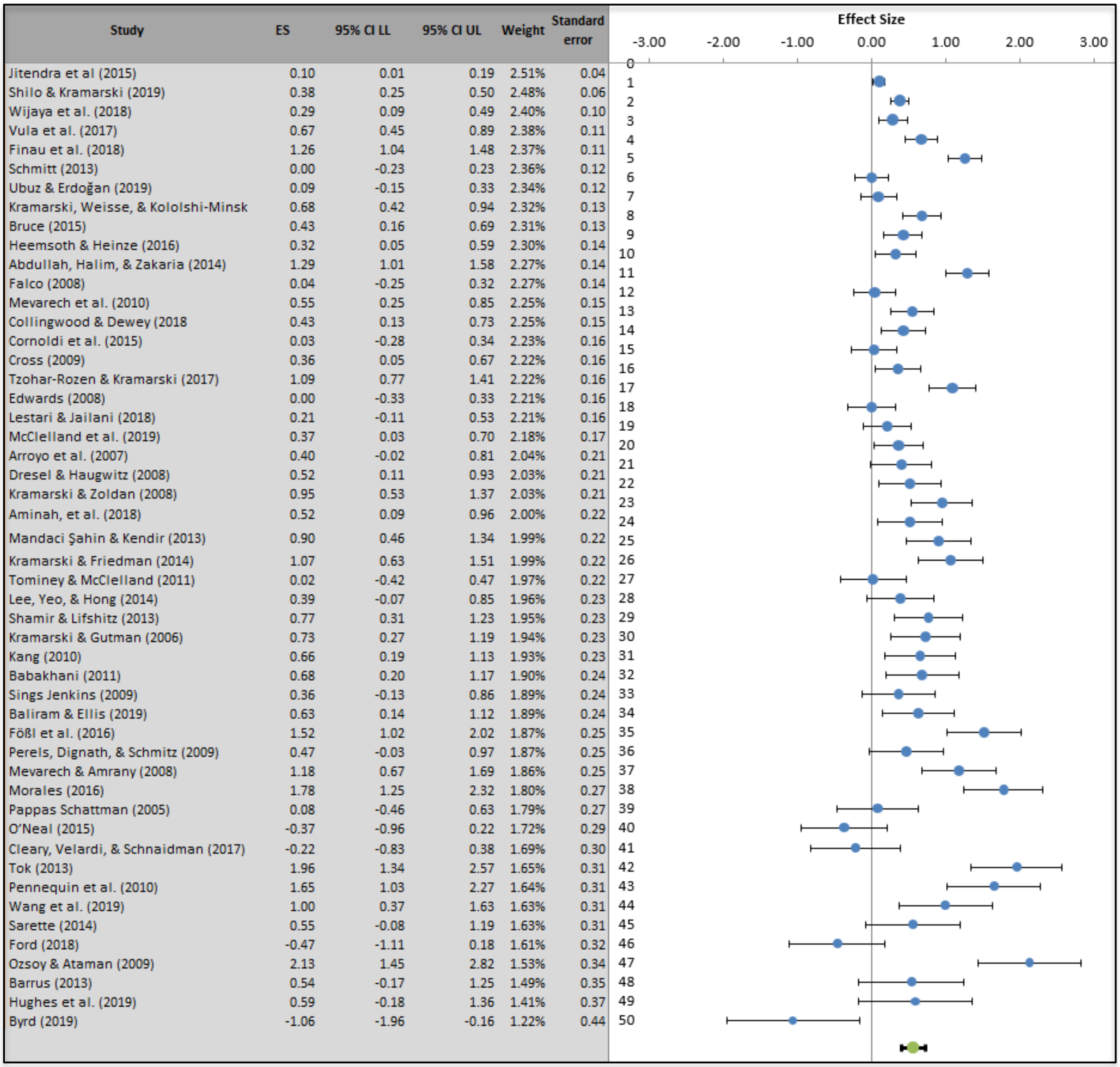
As shown in Table 27, the Q-statistic of 439.77 indicates the total amount of variation from the mean in the meta-analysis, calculated by summing the weighted, squared differences (or Weighted Sum of Squares) from the mean of the individual studies (Borenstein et al., 2009, p. 109). Although this value is high, it is partly due to the high number of individual studies included in the review, and is therefore not as relevant a marker of heterogeneity as I^2, which is a measure of the total variation in effects minus the variation expected given the degrees of freedom (i.e., the number of included studies minus one). This residual variation represents the proportion of the variation assumed to be real, or not due to sampling error. In this case, I^2 is reported as 86.58%, meaning that nearly all of the variation in the individual study effects is assumed to be real. The other values reported for heterogeneity are T^2=0.17 and T=0.41. These values can be interpreted as the variance and the standard deviation of the "true" effects, respectively (Borenstein et al., 2009, p. 111). Higher markers of heterogeneity could indicate important disparities between the included studies, which will be explored further below. First, however, it was thought to be beneficial to consider an alternate approach to generating the effect sizes from the original studies, which could also impact the degree of homogeneity found in the meta-analysis.

Q	439.77
p_q	0.000
I^2	86.57%
T^2	0.17
T	0.41

Table 27. Markers of heterogeneity for the posttest-only meta-analysis.

4.2.4 Pre/post and "best guess" meta-analysis results

As explained in detail in the previous chapter, I sought to incorporate all relevant data from the primary studies, and to adjust for potential baseline differences. Therefore, I performed a pre/post meta-analysis using a random-effects model and included 50 studies out of the original 60 from the review set. The remaining 10 studies either did not have a pretest comparable to the posttest, did not report it, or reported it such that MC/SRL and non-MC/SRL groups could not be compared. The results of the pre/post meta-analysis can be seen in Figure 20 and Table 28. The effect size is now *d*=0.56, SE=0.08, with a 95% confidence interval of 0.40 to 0.73. Regarding heterogeneity markers, the total variance is slightly less in the pre/post meta-analysis than in the posttest-only meta-analysis (Q=404.49, 439.77, respectively), but the proportion of “real” variance in the effects is slightly higher for the pre/post meta-analysis, with I²=87.89% compared to I²=86.58% for the posttest-only meta-analysis.



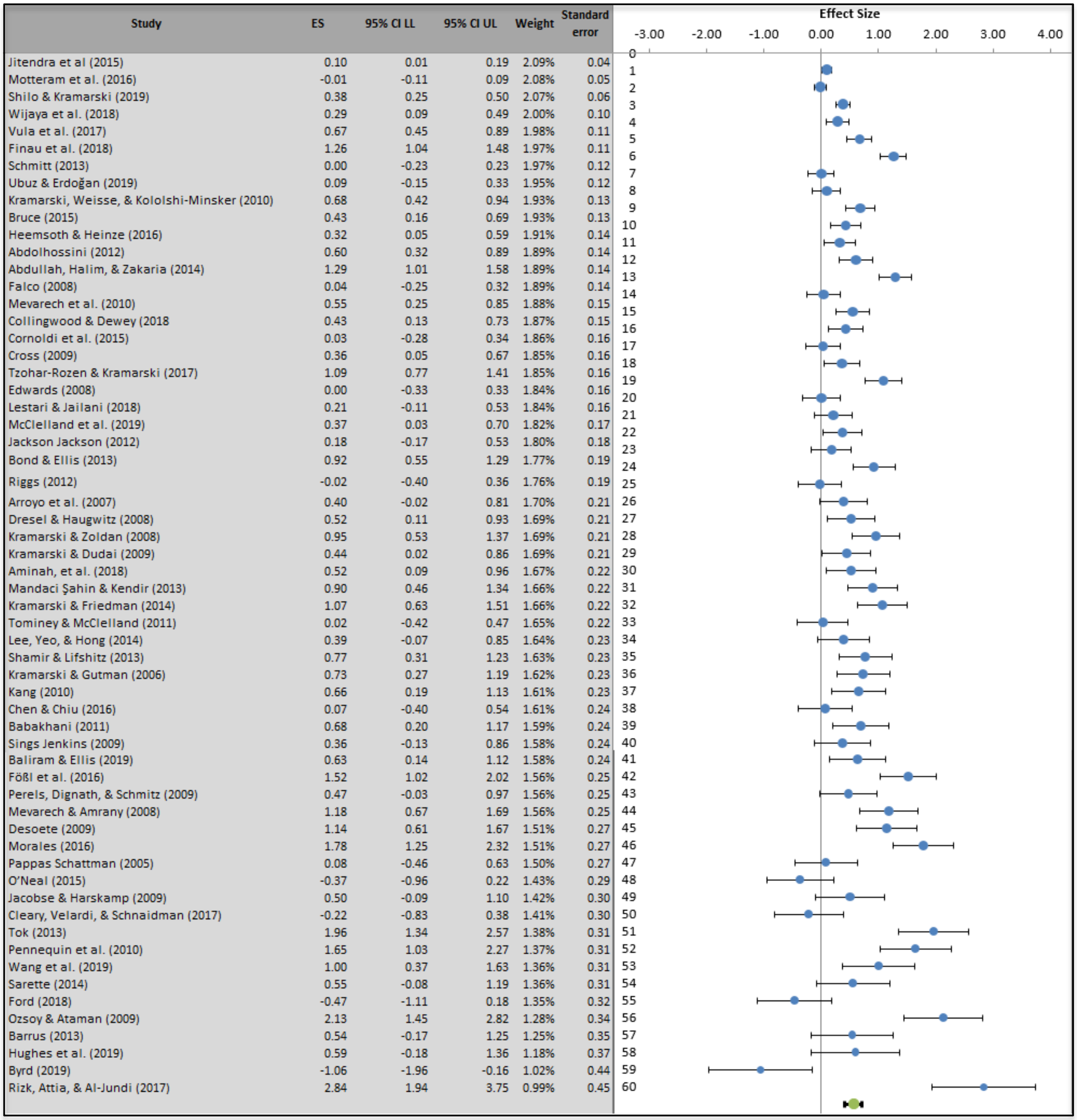
Study	ES	95% CI LL	95% CI UL	Weight	Standard error
Jitendra et al (2015)	0.10	0.01	0.19	2.51%	0.04
Shilo & Kramarski (2019)	0.38	0.25	0.50	2.48%	0.06
Wijaya et al. (2018)	0.29	0.09	0.49	2.40%	0.10
Vula et al. (2017)	0.67	0.45	0.89	2.38%	0.11
Finau et al. (2018)	1.26	1.04	1.48	2.37%	0.11
Schmitt (2013)	0.00	-0.23	0.23	2.36%	0.12
Ubuz & Erdoğan (2019)	0.09	-0.15	0.33	2.34%	0.12
Kramarski, Weisse, & Kololshi-Minsk	0.68	0.42	0.94	2.32%	0.13
Bruce (2015)	0.43	0.16	0.69	2.31%	0.13
Heemsoth & Heinze (2016)	0.32	0.05	0.59	2.30%	0.14
Abdullah, Halim, & Zakaria (2014)	1.29	1.01	1.58	2.27%	0.14
Falco (2008)	0.04	-0.25	0.32	2.27%	0.14
Mevarech et al. (2010)	0.55	0.25	0.85	2.25%	0.15
Collingwood & Dewey (2018	0.43	0.13	0.73	2.25%	0.15
Cornoldi et al. (2015)	0.03	-0.28	0.34	2.23%	0.16
Cross (2009)	0.36	0.05	0.67	2.22%	0.16
Tzohar-Rozen & Kramarski (2017)	1.09	0.77	1.41	2.22%	0.16
Edwards (2008)	0.00	-0.33	0.33	2.21%	0.16
Lestari & Jailani (2018)	0.21	-0.11	0.53	2.21%	0.16
McClelland et al. (2019)	0.37	0.03	0.70	2.18%	0.17
Arroyo et al. (2007)	0.40	-0.02	0.81	2.04%	0.21
Dresel & Haugwitz (2008)	0.52	0.11	0.93	2.03%	0.21
Kramarski & Zoldan (2008)	0.95	0.53	1.37	2.03%	0.21
Aminah, et al. (2018)	0.52	0.09	0.96	2.00%	0.22
Mandaci Şahin & Kendir (2013)	0.90	0.46	1.34	1.99%	0.22
Kramarski & Friedman (2014)	1.07	0.63	1.51	1.99%	0.22
Tominey & McClelland (2011)	0.02	-0.42	0.47	1.97%	0.22
Lee, Yeo, & Hong (2014)	0.39	-0.07	0.85	1.96%	0.23
Shamir & Lifshitz (2013)	0.77	0.31	1.23	1.95%	0.23
Kramarski & Gutman (2006)	0.73	0.27	1.19	1.94%	0.23
Kang (2010)	0.66	0.19	1.13	1.93%	0.23
Babakhani (2011)	0.68	0.20	1.17	1.90%	0.24
Sings Jenkins (2009)	0.36	-0.13	0.86	1.89%	0.24
Baliram & Ellis (2019)	0.63	0.14	1.12	1.89%	0.24
Fößl et al. (2016)	1.52	1.02	2.02	1.87%	0.25
Perels, Dignath, & Schmitz (2009)	0.47	-0.03	0.97	1.87%	0.25
Mevarech & Amrany (2008)	1.18	0.67	1.69	1.86%	0.25
Morales (2016)	1.78	1.25	2.32	1.80%	0.27
Pappas Schattman (2005)	0.08	-0.46	0.63	1.79%	0.27
O’Neal (2015)	-0.37	-0.96	0.22	1.72%	0.29
Cleary, Velardi, & Schnaidman (2017)	-0.22	-0.83	0.38	1.69%	0.30
Tok (2013)	1.96	1.34	2.57	1.65%	0.31
Pennequin et al. (2010)	1.65	1.03	2.27	1.64%	0.31
Wang et al. (2019)	1.00	0.37	1.63	1.63%	0.31
Sarette (2014)	0.55	-0.08	1.19	1.63%	0.31
Ford (2018)	-0.47	-1.11	0.18	1.61%	0.32
Ozsoy & Ataman (2009)	2.13	1.45	2.82	1.53%	0.34
Barrus (2013)	0.54	-0.17	1.25	1.49%	0.35
Hughes et al. (2019)	0.59	-0.18	1.36	1.41%	0.37
Byrd (2019)	-1.06	-1.96	-0.16	1.22%	0.44

Figure 20. Pre/post MA forest plot, studies ordered by standard error.

Result	**Pre/post MA**	**“Best guess” MA**	**Posttest-only MA**
Effect size	0.56	0.56	0.46
Standard error	0.08	0.08	0.08
95% confidence interval	0.40 to 0.73	0.41 to 0.71	0.30 to 0.62
Q	404.49	520.55	439.28
p_q	0.000	0.000	0.000
I^2	87.89%	88.67%	86.57%
T^2	0.18	0.18	0.17
T	0.42	0.42	0.41

Table 28. Results for pre/post, “best guess,” and posttest-only meta-analyses.

Since the pre/post meta-analysis contains fewer study-level effects than the original meta-analysis, it was decided to add back the 10 excluded studies to produce a “best guess” meta-analysis. Although these 10 studies could only be included using the posttest-only effect sizes, adding them back was expected to increase the clarity of the effects and reduce heterogeneity. The results are shown in Figure 21 and the second column of Table 28. Unexpectedly, the “best guess” meta-analysis produced an essentially equivalent effect size to the pre/post meta-analysis, and it increased the total variance (Q=520.55), while showing the largest proportion of “real” variance between effects, I^2=88.67%. As different procedures were utilised to produce the effect sizes in this meta-analysis, this may have resulted in the increased heterogeneity. Based on heterogeneity markers, it seems that the pre/post and “best guess” meta-analysis are not superior to the original posttest-only meta-analysis, which has the advantage of greater conceptual similarity between the effects. Because the posttest-only meta-analysis incorporated more consistent effects from a greater number of studies, the decision was made to revert to the original, posttest-only meta-analysis for the remainder of the quantitative investigation. Another reason to use the posttest-only result is that this is a more conservative estimate of the combined effect (ES=0.46 vs. 0.56 pre/post), and this might better predict applications of the findings by teachers.



Study	ES	95% CI LL	95% CI UL	Weight	Standard error
Jitendra et al (2015)	0.10	0.01	0.19	2.09%	0.04
Motteram et al. (2016)	-0.01	-0.11	0.09	2.08%	0.05
Shilo & Kramarski (2019)	0.38	0.25	0.50	2.07%	0.06
Wijaya et al. (2018)	0.29	0.09	0.49	2.00%	0.10
Vula et al. (2017)	0.67	0.45	0.89	1.98%	0.11
Finau et al. (2018)	1.26	1.04	1.48	1.97%	0.11
Schmitt (2013)	0.00	-0.23	0.23	1.97%	0.12
Ubuz & Erdoğan (2019)	0.09	-0.15	0.33	1.95%	0.12
Kramarski, Weisse, & Kololshi-Minsker (2010)	0.68	0.42	0.94	1.93%	0.13
Bruce (2015)	0.43	0.16	0.69	1.93%	0.13
Heemsoth & Heinze (2016)	0.32	0.05	0.59	1.91%	0.14
Abdolhossini (2012)	0.60	0.32	0.89	1.89%	0.14
Abdullah, Halim, & Zakaria (2014)	1.29	1.01	1.58	1.89%	0.14
Falco (2008)	0.04	-0.25	0.32	1.89%	0.14
Mevarech et al. (2010)	0.55	0.25	0.85	1.88%	0.15
Collingwood & Dewey (2018	0.43	0.13	0.73	1.87%	0.15
Cornoldi et al. (2015)	0.03	-0.28	0.34	1.86%	0.16
Cross (2009)	0.36	0.05	0.67	1.85%	0.16
Tzohar-Rozen & Kramarski (2017)	1.09	0.77	1.41	1.85%	0.16
Edwards (2008)	0.00	-0.33	0.33	1.84%	0.16
Lestari & Jailani (2018)	0.21	-0.11	0.53	1.84%	0.16
McClelland et al. (2019)	0.37	0.03	0.70	1.82%	0.17
Jackson Jackson (2012)	0.18	-0.17	0.53	1.80%	0.18
Bond & Ellis (2013)	0.92	0.55	1.29	1.77%	0.19
Riggs (2012)	-0.02	-0.40	0.36	1.76%	0.19
Arroyo et al. (2007)	0.40	-0.02	0.81	1.70%	0.21
Dresel & Haugwitz (2008)	0.52	0.11	0.93	1.69%	0.21
Kramarski & Zoldan (2008)	0.95	0.53	1.37	1.69%	0.21
Kramarski & Dudai (2009)	0.44	0.02	0.86	1.69%	0.21
Aminah, et al. (2018)	0.52	0.09	0.96	1.67%	0.22
Mandaci Şahin & Kendir (2013)	0.90	0.46	1.34	1.66%	0.22
Kramarski & Friedman (2014)	1.07	0.63	1.51	1.66%	0.22
Tominey & McClelland (2011)	0.02	-0.42	0.47	1.65%	0.22
Lee, Yeo, & Hong (2014)	0.39	-0.07	0.85	1.64%	0.23
Shamir & Lifshitz (2013)	0.77	0.31	1.23	1.63%	0.23
Kramarski & Gutman (2006)	0.73	0.27	1.19	1.62%	0.23
Kang (2010)	0.66	0.19	1.13	1.61%	0.23
Chen & Chiu (2016)	0.07	-0.40	0.54	1.61%	0.24
Babakhani (2011)	0.68	0.20	1.17	1.59%	0.24
Sings Jenkins (2009)	0.36	-0.13	0.86	1.58%	0.24
Baliram & Ellis (2019)	0.63	0.14	1.12	1.58%	0.24
Fößl et al. (2016)	1.52	1.02	2.02	1.56%	0.25
Perels, Dignath, & Schmitz (2009)	0.47	-0.03	0.97	1.56%	0.25
Mevarech & Amrany (2008)	1.18	0.67	1.69	1.56%	0.25
Desoete (2009)	1.14	0.61	1.67	1.51%	0.27
Morales (2016)	1.78	1.25	2.32	1.51%	0.27
Pappas Schattman (2005)	0.08	-0.46	0.63	1.50%	0.27
O'Neal (2015)	-0.37	-0.96	0.22	1.43%	0.29
Jacobse & Harskamp (2009)	0.50	-0.09	1.10	1.42%	0.30
Cleary, Velardi, & Schnaidman (2017)	-0.22	-0.83	0.38	1.41%	0.30
Tok (2013)	1.96	1.34	2.57	1.38%	0.31
Pennequin et al. (2010)	1.65	1.03	2.27	1.37%	0.31
Wang et al. (2019)	1.00	0.37	1.63	1.36%	0.31
Sarette (2014)	0.55	-0.08	1.19	1.36%	0.31
Ford (2018)	-0.47	-1.11	0.18	1.35%	0.32
Ozsoy & Ataman (2009)	2.13	1.45	2.82	1.28%	0.34
Barrus (2013)	0.54	-0.17	1.25	1.25%	0.35
Hughes et al. (2019)	0.59	-0.18	1.36	1.18%	0.37
Byrd (2019)	-1.06	-1.96	-0.16	1.02%	0.44
Rizk, Attia, & Al-Jundi (2017)	2.84	1.94	3.75	0.99%	0.45

Figure 21. "Best guess" meta-analysis forest-plot, studies ordered by standard error.

4.3 Moderator and subgroup analyses results

To address the second research question, and based on the narrative synthesis and previous research, I considered several variables for a possible exploratory analysis. Although various qualitative aspects of the included studies were examined, many of these were reported inconsistently or were only included in a minority of studies. It was considered preferable to choose more clearly defined study aspects for exploratory analysis, since these results would potentially lead to more robust findings. The term “exploratory” is used here to indicate that the chosen variables were not part of the original design for the meta-analysis and did not impact the protocol for the systematic search. Because the searching and screening process was focused on generating an estimate of effect for all studies under the MC/SRL umbrella, this constraint may make the body of studies in the final review unsuitable to reliably estimate the effects of specific subgroups or moderators. The intention instead was to identify fruitful areas for further study, and to outline foci for potential confirmatory syntheses. I also sought to avoid over-analysing the data, leading to potentially spurious results. Thus, only a few subgroups and moderators are reported on from many possible options. The rationale for each is presented here, along with a brief interpretation of the results. Further discussion of the exploratory analyses follows in the next chapter.

4.3.1 Structured problem-solving with metacognitive self-questioning

Regarding the activities and strategies used in the interventions, studies frequently referred to IMPROVE (Mevarech & Kramarski, 1997), which uses specific prompts or self-questions to guide learners through a structured problem-solving approach[33] (see Table 29). This review features several studies led by the developers of IMPROVE, along with others that used the approach or referred to it as inspiration for a newly developed intervention. In addition, some studies did not refer to IMPROVE but to other structured problem-solving approaches utilising metacognitive questions or prompts, such as one based on Pólya (1971, cited in Lee et al., 2014, p. 466). I decided to examine the possible impact of including IMPROVE or a similar approach on the mathematics outcomes. To do this, studies were coded based on the extracted qualitative data, as either IMPROVE or a similar intervention, IMPROVE plus other main elements (e.g., computerised problem-solving and feedback, reflective writing, affective regulation), or not based on IMPROVE or a similar approach. Using these categories, it was found that only four reports used IMPROVE alone, so I combined the two IMPROVE categories. Ultimately, 28 studies were classified as having IMPROVE or a similar approach as a main intervention component, while 32 studies were not. The latter group included play- and game-based, reflective writing, e-book, goal-setting, calibration, motivational, and “study skills” interventions, as well as those that did train problem-solving strategies but did not expect

[33] Note that the authors consider all these elements as comprising three overall components: “metacognitive questioning, cooperative learning, and systematic provision of feedback-corrective-enrichment” (pp. 373-374). Not all studies by these researchers and others that cite IMPROVE have used the same elements as the original study, but metacognitive questioning has been the most commonly used element.

students to apply them using a set procedure for every problem. Because of this, the group of interventions coded as IMPROVE or similar may represent a more cohesive group than the "not IMPROVE" category.

Programme element	**Explanation**
Mnemonic device	Introducing new concepts Metacognitive questioning Practising Reviewing and reducing difficulties Obtaining mastery Verification Enrichment (p. 369)
Metacognitive questions	Pupils use index cards with metacognitive questions to prompt themselves during different problem-solving stages of individual practice and group work. Problem types include: Comprehension questions, for understanding the problem. Strategic questions, related to problem-solving approaches. Connection questions, for drawing comparisons to other types of tasks.
Cooperative group work	Pupils work through challenging mathematical tasks in mixed-ability groups and make use of their different forms of "prior knowledge," as they discuss, question, suggest, challenge and explain approaches to solving them.
Formative assessment, feedback	Pupils complete a unit test every 10 lessons, designed to probe higher-level thinking and application, not only mathematics skills. Low performers get feedback, work together to complete "corrective activities," and then take a parallel form of the test.
Enrichment	Higher performers on formative tests complete more difficult "enrichment" activities to build mathematical reasoning instead of remedial work. They work with similarly able students.

Table 29. Original IMPROVE programme elements, from Mevarech and Kramarski (1997).

Using the posttest-only effect sizes, a subgroup analysis was run in Meta-Essentials based on a random-effects model with T (tau) separated by subgroups. This assumes the true effect is more similar within each subgroup than over the whole study set. Based on the subgroup analysis, it might be interpreted that there was a small additional benefit to including structured problem-solving with metacognitive prompts or self-questions within the MC/SRL training. The overall ES of such studies was reported as 0.53 (95% CI=0.30 to 0.77) while those not coded as IMPROVE or similar were reported as having a summary estimate of ES=0.40 (CI=0.17 to 0.64). The combined ES of the studies with IMPROVE or similar was also higher than that of the overall posttest only meta-analysis (ES=0.46, 95% CI=0.30 to 0.63). However, similar to the overall meta-analysis, there is a high degree of reported heterogeneity within each subgroup, as seen in Table 30. With I² values around 86-87%, there is substantial "real" variance within each subgroup. In fact, Meta-Essentials reports a pseudo R^2 for this subgroup analysis of 1.01%, meaning the two categories explain almost none of the total variance in effects. Based on the estimates of variance/heterogeneity (T² and T), there is a slightly wider range of potential "true" effects for studies without an IMPROVE-like component than for those with one.

Subgroup	Studies	Sub-group effect size	95% Confidence interval	Q	p_q	I²	T²	T
IMPROVE or Similar	28	0.53	0.29 to 0.76	197.92	0.000	86.36%	0.15	0.39
Not IMPROVE or similar	32	0.40	0.17 to 0.64	241.19	0.000	87.15%	0.22	0.46

Table 30. Subgroup analysis for IMPROVE and non-IMPROVE studies.

4.3.2 "Dose" of MC/SRL training

I also explored whether there could be differential outcomes from the MC/SRL interventions based on the "dose" of the training. Previous reviews have considered this question (e.g., de Boer, Donker, & van der Werf, 2014; Dignath & Büttner, 2008; Wang & Sperling, 2020), with mixed results. The rationale is that a longer intervention would allow students to build more extensive knowledge and habits related to MC/SRL practices that could be beneficial for learning and performance. "Dose" in this case was operationalised as the time in weeks from the start to the end of the intervention, regardless of how many intervention sessions were held during each week and the length of the sessions. This choice was made based on the variability in reporting about the sessions. In some cases, the session length is not clearly stated, or it is unclear how much of each session was devoted to the MC/SRL training or to the standard mathematical teaching. In addition, were total hours of the intervention used as a moderator, this would obscure the potential effects of teachers' spontaneous reinforcement of the MC/SRL training throughout the week, not only during scheduled sessions.[34] Having a longer intervention that is more spaced out could therefore have a higher effect on students'

[34] Whether or not teachers did spontaneously reinforce the MC/SRL training outside scheduled sessions is generally not reported.

learning and performance than one with the same number of total training hours compressed into a shorter period of time.

To explore “dose” as a moderator, total weeks of MC/SRL training was entered into the Meta-Essentials workbook. Wherever this was ambiguous, an estimate was made based on the description in the report. For example, if the authors stated that the intervention was about one month, then it was coded as four weeks. One semester was coded as 18 weeks. The shortest interventions were one week or less (coded as one week) and the longest was 40 weeks. For this analysis, the effect size used was the posttest-only effect size, computed using a random effects model. Meta-Essentials performs a regression using only one moderator at a time, in this case, intervention weeks. The results show that there is no meaningful impact on the effect size based on the length of the intervention as coded with R²=0.19% (see Figure 22). This means that almost none of the variance between the effect sizes in the primary studies can be explained by the length of the intervention as coded.

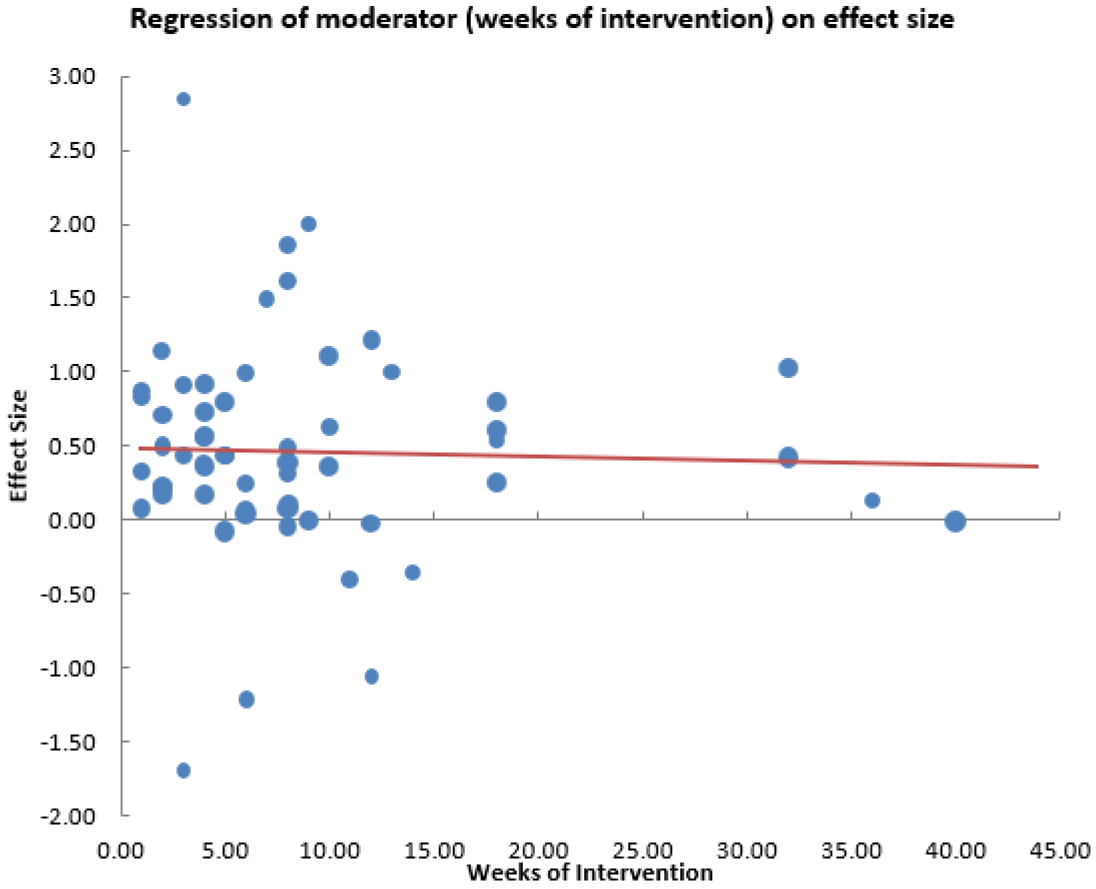


Figure 22. *“Dose” moderator analysis for posttest-only MA.*
Total length of intervention in weeks was the “dose” marker.

4.3.3 Participant age

It was further considered whether the effects of the interventions might vary depending on the ages of the pupils involved in the studies. Undoubtedly, developmental stage plays a role in MC/SRL functioning (Siegler & Chen, 1998; Waters & Kunnmann, 2010), so the effectiveness of MC/SRL training could vary based on participants' ages. For example, pupils might need to possess basic language skills, emotional self-awareness, working memory, time perception, and an understanding of cause and effect in order to reflect on their learning, articulate and compare strategies, and engage in planning. In addition, a certain level of domain-based knowledge could be required in order to enable students to be strategic in their mathematics learning and performance. Although many of the included studies reported details about students' abilities and mathematics knowledge, it was decided that ability-based comparisons would not be appropriate. For example, some studies reported having "inclusive" or multi-ability classrooms, while others included only higher- or lower-ability students. Because the review included studies from many educational contexts, it is doubtful whether such classifications would be consistent from study to study. However, age-based comparisons would be fairer. The current review included by design all studies with pupils in the years of general and/or compulsory education, roughly corresponding to ages three to 18, though not all compulsory education systems include all these ages. The goal was to determine the extent to which a similar effect could be found across different ages. Age was coded based on the mean age of participants, if reported, or based on the grade or year-level if actual ages were not reported. Age reported was always used, even if it did not match expectations based on the grade or year of school. Table 31 shows the coding whenever the terms "grade" or "year" were used, based on the ages commonly included in US and UK schools, respectively. It is acknowledged that the actual ages of students could vary by a year or more. Whenever multiple ages or grades were reported, an effort was made to determine average age based on the number of students of each age or grade involved in the study. Because age seemed to be age at pretest for most of the reports, when both pre- and posttest ages were given, the former were used. Only one average age was entered for each study. Due to reporting differences, it was not possible to consider different ages within or between treatment groups or teaching units for this analysis.

US grade level	UK school year	Age coded
Preschool	Reception	4
Kindergarten	1	5
1st	2	6
2nd	3	7
3rd	4	8
4th	5	9
5th	6	10
6th	7	11
7th	8	12
8th	9	13
9th	10	14
10th	11	15
11th	12	16
12th	13	17

Table 31. Age coded based on US grade or UK year level.

Average age rounded to three decimal places was entered as a moderator into the Meta-Essentials workbook for the posttest-only, random effects meta-analysis. Results show that age is not a reliable predictor of effect size, with R^2=0.93% (see Figure 23). This means that almost none of the variance in the effects can be explained by the participants' ages as coded. The Figure 23 shows that there is a great variability in the effects across the age range, while interventions done with the youngest children seems to have more consistent effects. This could be a spurious finding, however, as most of the studies with very young children used the same intervention, the Red Light, Purple Light circle time games (Tominey & McClelland, 2011; Schmitt, 2013; Schmitt et al., 2015; McClelland et al., 2019). Being coded as a younger age study could therefore be a proxy for MC/SRL interventions that are more play-based. Most average ages of the included studies ranged between eight and 15, and there were a variety of interventions and effect sizes within this range. Fewer studies were done with students older than 15, and more negative effects were seen in this age range than in the other ages.

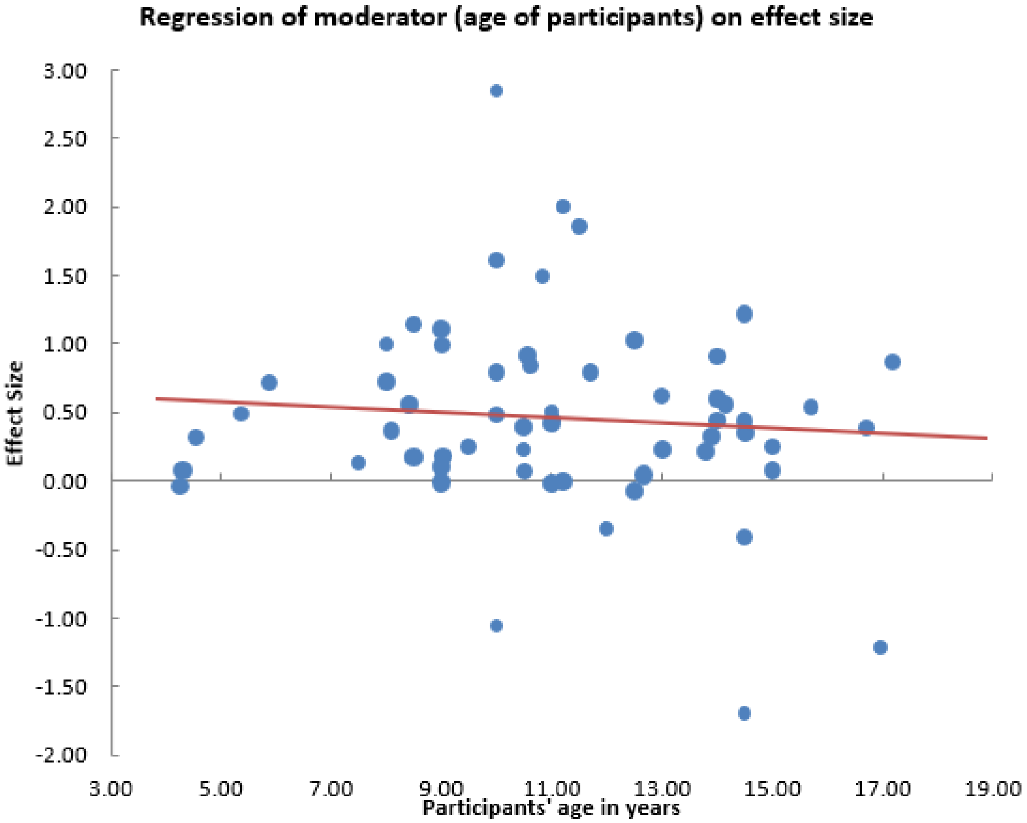


Figure 23. Moderator analysis results for post-only effect sizes and age of participants.

4.3.4 Report type

Research reported in academic journals tends to show a more positive effect than unpublished or “grey” literature. Although the reasons may be complex, according to Song, Hooper, & Loke, (2013, p. 73), researchers have admitted they sometimes de-prioritise publishing results, presumably because they are lacklustre. The fact that null or negative findings may be less likely to be published is one reason for examining the risk of publication bias in a meta-analysis. As shown in section 4.2.2, I concluded there was little risk of overestimating the combined effect size due to leaving out studies with lower effects, but it was still possible that the different types of reports included would be linked to different effect sizes. As mentioned, this review included a higher proportion of dissertations and conference papers than seen in previous reviews, so different effects by publication type could have more of an impact overall. To explore this, I coded each study based on four report type categories: conference papers, dissertations or theses, journal articles, and technical reports. The only technical report included was Motteram et al. (2016), which had a very large sample size (n = 1507) but an overall null result (ES = -0.01), and including this as a “subgroup” produced an error in the Meta-Essentials programme due to unmet assumptions. I decided to exclude Motteram et al. (2016) from this subgroup analysis for this reason. The subgroup analysis used the posttest-only effect sizes and was based on an RE model with T separated by subgroups.

The results of the report type subgroup analysis (see Table 32) show that this variable predicts more of the variance in effect than any other variable considered, with a pseudo R^2 of 13.22%. As predicted, the journal article subgroup (40 studies) showed the most positive results (ES=0.62), while conference papers (4 studies) had lower effects (ES=0.44), and dissertations and theses (15 studies) had negligible effects (ES=0.06). Recall that all of the dissertations and theses came from US-based researchers, while those in the other report categories had an extensive geographic and political range. This means that local cultural factors could also play a role in the different effects seen in the dissertation category. In addition, while dissertations had the lowest effects overall, they also showed the widest confidence intervals and largest values for T^2 and T, meaning the effects reported in dissertations are more inconsistent than those in other report types. On the other hand, the I^2 value for journal articles is slightly more than that for dissertations, implying there could be more “real” variance in effects within the former than the latter categories, but this may not be a meaningful finding since the journal article category contains more than twice the number of studies and has a higher Q value to begin with. The significance scores (p_q) are reported here for transparency, but due to the low number of studies and lack of statistical power in the conference paper category, it may not be meaningful to compare the different report type categories on this metric. Overall, it is clear there is a difference in the effects between these subgroups (Q between/model = 12.74), while the variation within each group is still much greater (Q within/residual = 83.64). This confirms the choice of an RE model for this analysis, and it also implies that the greater part of the variation in effects is related to other, still undefined factors, rather than being linked to the type of report. This issue is considered at more length in the discussion chapter following.

Report type	**Studies**	**Sub-group effect size (SE)**	**95% Confidence interval**	**Q**	**p_q**	**I^2**	**T^2**	**T**
Conference papers	4	0.44 (0.18)	0.09 to 0.79	3.76	0.288	20.23%	0.01	0.10
Dissertations / theses	15	0.06 (0.20)	-0.34 to 0.46	77.13	0.000	81.85%	0.22	0.47
Journal articles	40	0.62 (0.09)	0.44 to 0.81	300.36	0.000	87.02%	0.18	0.42

Table 32. Report type subgroup analysis based on posttest-only effects.
An RE model with T separate for subgroups was used. One technical report is excluded.

Chapter 5: Discussion and Conclusions

This review has focused on two key areas for education practice, mathematics learning and the development of self-regulation and metacognitive knowledge and skills. Unlike the acquisition of natural languages, for which humans seem to have a natural affinity the development of mathematics skills is sometimes consciously effortful (Gafoor & Sarabi, 2015), with inadequate integration and utilisation of knowledge and strategies (Garafalo, 1989, p. 503). Due to its incremental nature, mathematics expertise can take many years to develop, but it can prove extremely valuable for individuals in both developing and developed nations (Schleicher, n.d.). In addition to the practical benefits of basic numeracy, such as being able to manage property, avoid financial exploitation, and self-advocate as a business-owner or employee, there are numerous avenues for personal advancement which have mathematics skill or achievement as a known “gatekeeper” (Douglas & Attewell, 2017). Respected and highly-compensated STEM professions, for example, may remain relatively closed to individuals from less-privileged backgrounds if they have inferior mathematics-related opportunities and outcomes (Kotok, 2017). Beyond the instrumental rationale, mathematics can be a powerful tool for understanding and communicating about the world in creative ways, and such a focus may improve the mathematics engagement of learners from all backgrounds. Focusing on achievement more indirectly, while emphasising the development of mathematics-related identities and self-efficacy, may also offer relief for students experiencing negative affect in the mathematics classroom. With these goals in mind, there is clear potential in metacognitive and self-regulated learning approaches. By building learners’ self-awareness and agency regarding their own learning, these approaches can help them adapt to the demands of different tasks and domains. In addition, MC/SRL programmes have shown benefits for achievement outcomes, particularly in mathematics (e.g., Dignath & Büttner, 2008; Higgins et al., 2005), and have been listed as among the most cost-effective of all interventions, for example in the EEF Teaching and Learning Toolkit (2021). Previous syntheses, as presented in Chapter 2 of this thesis, have found medium to high effects for MC/SRL programmes within different domains, school-levels, and national contexts, but there has been no recent meta-analysis focusing on MC/SRL impacts on mathematics outcomes specifically, nor a consensus on the essential intervention elements to produce such effects. The current research has sought to fill this gap in knowledge.

5.1 Chapter outline

This systematic review and meta-analysis investigated the effects of MC/SRL-based interventions on mathematics outcomes for school-aged pupils since 2005. A second goal was to understand how differences in effects could result from specific intervention elements or design- or implementation-related factors. This information was sought to assist researchers, practitioners, and other educational stakeholders in making decisions about which programmes to implement in their specific contexts and in conducting future empirical and theoretical work in the MC/SRL area. This closing chapter offers a discussion of the results of the current synthesis, starting with the meta-analysis and the moderators considered and following with the narrative synthesis. Overall, this review confirmed results of previous studies showing MC/SRL programmes have great potential to benefit teaching and learning, but there are nuances and uncertainties remaining around this message. I explain how the combined effect size could have different meanings in various schooling contexts, as well as how the findings about moderators could be applied. I next present some potential limitations of the current research as I reflect on the included studies, the review methods, and the chosen research foci. As mentioned, reporting issues somewhat obscured answers to my research questions, but several of my choices for the review could also come under scrutiny, especially if a replication or update were planned. In the following section, I outline several recommendations for future research and practice. With regard to the former, I argue for consideration of how reporting standards could and should be applied to quasi-experimental, observational, and other non-RCT designs. I also suggest more consistent use of validated, broader-scope assessments. Both of these would strengthen the evidence around educational interventions, as well as making them more open to comparison, replication, application, and synthesis. Within the area of metacognition and self-regulated learning, I recommend more work to clarify “what counts” as an MC/SRL programme and to link this to specific features that school leaders, teachers, and researchers can observe and implement. Such features can then be used to categorise and compare MC/SRL programmes more productively. With regard to practice, my recommendation is that MC/SRL implementation be viewed as a worthwhile endeavour, likely to produce gains with sufficient training, support, and time. Near-term outcomes may be seen with structured problem-solving approaches that prompt learners to self-question, as these may spur better application of mathematics knowledge. MC/SRL training should not stop there, however, as a more comprehensive approach, one that incorporates affective, behavioural, and motivational components and ensures opportunities to practise meta-level skills, may be more likely to empower students to learn, grow, and achieve throughout their lives.

5.2 Meta-analysis findings

I conducted a systematic search and screening process, followed by structured data extraction, and a meta-analysis of study-level effects to address the first research question for this review (shown in Appendix 2): *During the last 15 years, what has been the effect of interventions based on theories of metacognition or self-regulation on the mathematics achievement/proficiency of school-aged learners?* I used posttest data on mathematics measures from included studies at the first assessment point following the end of the MC/SRL training period. All MC/SRL-based participant groups were combined and compared with all non-MC/SRL groups, to generate a single effect size from each study, and these were incorporated with a weighted, RE meta-analysis in Meta-Essentials. The resulting combined effect from 60 included studies was 0.46 (SE=0.08), with a 95% confidence interval between 0.30 and 0.62. This combined effect demonstrates a benefit for MC/SRL programmes similar to that found in previous reviews, which ranged from ES=0.54 to ES=0.86 for all subjects and ES=0.23[35] and ES=1.01 in mathematics.[36] The current estimate is somewhat more conservative, as discussed further below.

5.2.1 The combined effect

The meaning of the 0.46 effect size from this systematic review needs further interpretation. It would likely be viewed as a "medium effect" even though it falls just shy of the 0.50 cut-off proposed by Cohen (1992, p. 156) for an effect ". . . likely to be visible to the naked eye of a careful observer." Cohen (1992) also noted that such an effect aligned with the average intervention effect reported in various fields, and this has been confirmed by Hattie's "Visible Learning research," which finds an average effect of 0.4 ("The Visible Learning Research," n.d.; Hattie, Biggs, & Purdie, 1996). Although an "average effect" may not seem remarkable, another way to interpret effect sizes is to translate them into "months of progress," following the lead of the EEF Teaching and Learning Toolkit. With the assumption that one year of schooling equates to one standard deviation of gain on standardised assessments, the EEF (2018, p. 28) suggests that an effect of 0.46 equates to approximately six months of schooling and represents a "high"

[35] This lower combined effect is the secondary school mathematics outcome from Dignath & Büttner (2008), but the primary level effect from that review is considerably higher (ES=0.96). They do not report a combined mathematics effect for all years.

[36] Here, I omit results from de Boer et al. (2018), which are the posttest to follow-up effects. That is, they show how much additional effect is correlated with a delayed assessment.

effect. The EEF's own research has led to a similar estimate to the current review, an additional seven months of progress for MC/SRL-type interventions across all pupils and domains. On the other hand, the EEF (2018, pp. 26-27) acknowledges some imprecision in these estimates, since the average yearly gain tends to decrease in the higher grades. This means a consistent effect of 0.46 would have a greater impact for older than for younger students. How the assessments were scaled to the intervention periods could also have an influence, as I discuss further below, and rate of normal learning may fluctuate across an academic year or term (Baird & Pane, 2019, p. 225). Using such estimates, there are also concerns that research users may erroneously conclude that interventions obviate the need for expert teaching and sufficient resources, leaving those susceptible to budget cuts. Baird & Pane (2019, p. 226) also caution that statistical uncertainty increases when converting effect sizes to other metrics. They recommend instead the use of "percentiles translation," or reporting the average change in percentile of median-level students, though they state this can also be calculated for other baseline points (p. 222). The advantages are that the measure is intuitive and common to educational contexts with fewer assumptions than other methods, such as using thresholds and benchmarks (i.e., criterion-referenced conversions). Yet there are some potential downsides to this, not covered by Baird and Pane (2019). The percentile method still assumes use of large-scale, standardised assessments, which do not always discriminate well for scorers at the ends of the distributions, as discussed by Cole et al. (2011).

Based on an analysis of 1,942 effect sizes from 747 randomised controlled trials of educational programmes, Kraft (2020) proposes new "benchmarks" for the interpretation of effect sizes: under 0.05 for a "small" effect, 0.05 to less than 0.20 for a "medium" effect, and 0.20 or greater for a "large" effect. Thus, the current and previous MC/SRL reviews would be interpreted as finding large effects. Only 18 out of 60 studies from my review fell below the "large" cutoff in Kraft's (2020) "schema." However, the appropriateness of judging the review by this schema may be uncertain since my included studies were not all randomised at the student level and did not all utilise standardised assessments. Kraft (2020) also discusses numerous factors that may be associated with higher or lower effects, such as the type and timing of assessments (p. 244), sampling and tailoring of the intervention to pupils' needs (p. 244-245), the standard deviation or variance measure chosen as the effect size unit (p. 245), the nature of the comparison group (p. 245), and differences in participation levels and attrition (p. 245-246). These variables could obscure "real" effects and create challenges for research synthesis, as found in the current review.

Kraft (2020) recommends “adapting” the benchmarks based on several considerations. First, according to Kraft (2020) effects are higher in language arts, especially during the early stages of literacy development, while mathematics effects are more consistent across grades 1 through 12, hovering around a “medium” effect of ES= 0.05 (p. 249). This confirms that the current 0.46 effect size in mathematics is large for every school level and may not need parsing based on different ages or years. Next, Kraft shows that larger samples and “broad achievement measures” (p. 248) are linked to lower effect sizes. In the current review, 50 out of 60 studies included 200 or fewer pupils. Assessments varied, with most studies reporting some type of validation process, but it is not clear if these would be considered broad measures, so could be a risk of effect size inflation based on these two factors. In this review, several larger-scale studies (e.g., Jitendra et al., 2015; Motteram et al., 2016) had lower effects, while Shilo and Kramarski (2019) included 824 participants and achieved an effect size of ES=0.39 using eight open-ended items adapted from Israeli national assessments. Based on Kraft (2020), it is possible that this larger effect is partially a result of the focused assessment tool and would be smaller if a more comprehensive test had been used.

A final recommendation from Kraft (2020) is to interpret effect sizes in light of the cost and scalability, interpreting per pupil costs of less than 500 “2016 constant dollars” as low cost, between $500 and $4,000 as moderate, and $4,000 or higher as high-cost. With a year of school costing over $10,000 per student[37] (Kraft, 2020, p. 247), a low-cost intervention would be one using 1/20 of funding. Although studies in this review generally did not report costs in detail, most MC/SRL programmes would be low- to medium-cost based on using classroom discussion and simple materials like worksheets, meaning that the 0.46 combined effect confirms the cost effectiveness of MC/SRL interventions. The use of digital devices would add to costs if they were not already provided to pupils. Scalability estimation would include monetary costs as well as other demands on the intervention site, such as training and support for making changes to classroom and school learning culture.

The above discussion shows that the value of the combined effect in practice is not straightforward to determine. Judgement and a knowledge of local factors must be used to determine when and how to apply research to practise and what effect would be valuable within each context. In an era when students are assessed frequently and with major consequences, schools will need to carefully consider what intervention impacts to aim for and what

[37] Clearly, this estimate may not apply to all schools globally.

investments to make to reach them. Based on this review, MC/SRL training is a worthwhile option to consider to meet mathematics performance goals.

5.2.2 Potential moderators of effect

I found an overall positive effect on mathematics outcomes for MC/SRL training, based on a distribution of 60 primary studies with no clear imbalance or indication of bias, yet there were moderately high heterogeneity values resulting from the meta-analysis. As a reminder, the I^2 value was 86.57% from the posttest-only meta-analysis, and this was not reduced in a meta-analysis incorporating pretest data, where reported. Partly for this reason, I chose the posttest-only combined effect as the official one, but the issue of heterogeneity remained. Heterogeneity could indicate meaningful differences in how the studies were designed or implemented, or in the nature of the included MC/SRL programmes, which might be associated with different effects. In planning the systematic review, I anticipated this, and my second research question was: *What specific factors, if any, are correlated with higher effectiveness for [MC/SRL] interventions?* Thus, I performed several exploratory analyses to check whether categorising the interventions based on factors gleaned from the narrative synthesis would reduce heterogeneity and indicate more beneficial aspects of MC/SRL training. Here I discuss possible reasons why several of these moderator checks showed clearer results than the others.

First, the subgroup analysis showed higher effects for interventions based on structured problem-solving with metacognitive prompts or self-questions than for programmes not including this. Because nearly half of the included studies included or were inspired by the IMPROVE approach (Mevarech & Kramarski, 1997), it seemed profitable to consider if IMPROVE and similar programmes led to better achievement. I found an ES=0.13 difference, with IMPROVE-type interventions being more effective (ES=0.53) than others (ES=0.40). This difference is small, and both intervention categories are still effective. Interestingly, the pseudo R^2 value was 1.01%, showing that the variation in effect between these two categories was minor compared with the internal variance. This is a result that could be confirmed with a more focused synthesis of all IMPROVE studies, as I did not differentiate between different variations of IMPROVE, such as those using a computer application to prompt students (e.g., Kramarski & Gutman, 2006) or incorporating emotional regulation components (e.g., Tzohar-Rozen & Kramarski, 2017 & 2013). Still, it is worth examining why IMPROVE-type programmes might lead to better mathematics performance. IMPROVE uses a mnemonic device to allow students to acquire and remember the approach more readily. It also gives students concrete steps to follow while executing a solution strategy and prompts them to check their performance at multiple stages.

Most of the assessments in this review employ complex problems, in which students need to synthesise information from words, numbers, and images, and execute multiple operations to produce an answer, and IMPROVE or similar approaches seek to help students tackle such items in a consistent and thoughtful way. In addition, IMPROVE is a relatively simple approach that is straightforward for teachers to train in a limited timeframe, so it might be easier to implement than other MC/SRL programmes. Thus, using IMPROVE-like training would almost certainly lead to an advantage in performance measures. On the other hand, studies in this review used IMPROVE in numerous variations, often combining it with other elements, and there is no clear signal about which version is better. In addition, IMPROVE by itself is very task-focused and does not encourage students to reflect deeply on the mathematics domain or their identities as mathematicians. Students may also need to build their conditional metacognitive knowledge (Pintrich, 2002) to decide when it is worthwhile to implement a structured problem-solving approach or only certain aspects of it, such as in the context of a timed test. Finally, there is a possibility that the differences in effects seen here are also related to publication type, as the non-IMPROVE-type group included 12 theses and dissertations, while the IMPROVE-type group included only three. This type of report had a much lower average effect than journal articles and conference papers, and this is considered next.

Traditionally, there has been a perception that research published in peer-reviewed journals is of higher quality than "grey literature" studies, which include conference papers and theses/dissertations. Previous reviews of MC/SRL programmes included fewer unpublished reports, perhaps reflecting this belief, but this could also be due to the greater difficulty in accessing grey literature or the fact that dissertations are longer and less practical to review. For the current review, I did not assume such studies were of inherently lower quality, and I was able to access many of them in my electronic searches, so one-third of my inclusions were non-journal articles. Based on this higher-than-average proportion, I checked whether there were differences in effects, and I found a substantially higher effect for journal articles (ES=0.62) than for theses/dissertations (ES=0.06). Conference papers, of which there were only four, had a combined effect size of 0.44, while the subgroup analysis did not permit including the one technical report, Motteram et al. (2016).[38] Only three out of 40 journal articles reported negative effects, and these patterns bear out previous research (Song, Hooper, & Loke, 2013). Chong et al. (2016) found that, out of 1,052 studies featured at a conference on anaesthesia, those reporting positive results were 1.42 times as likely as those reporting negative results to be

[38] I calculated this study-level effect size as -0.01.

published in journals within 10 years. The reasons for these differences may be complex. Researchers may self-censor and refrain from releasing negative findings, or release them in a lower-access way, such as in a conference paper. They could also choose to report only "significant" outcomes of their studies, while holding back other results. Journal reviewers may also view negative results less favourably and not recommend articles for publication. Theses and dissertations should not be as prone to the "file drawer problem" (Rosenthal, 1979), since even with null or negative results they are still likely to be submitted and become searchable in a review.

It is also possible that there are real differences in effects, with the interventions reported in conference papers and dissertations having lower impacts on achievement. In the current review, sample sizes were not substantially larger in journal articles, but theses and dissertations still likely represent "smaller-scale studies" since they often featured teachers working with their own students, with possibly fewer resources and less expert oversight. In one case, Schmitt (2013), the author worked with a larger research team using the "Red Light, Purple Light" intervention, and this dissertation chapter, even though it featured a modest mathematics effect (ES=0.08), was later published as a journal article. In any case, it was important for the current review to include studies designed and implemented by teachers because these may better reflect outcomes that practitioners should anticipate when implementing an MC/SRL programme. For example, See, Gorard, and Siddiqui (2016) show that teachers' operationalisation of research-based concepts like "feedback" may differ from researchers'. In addition, teachers may find it time-consuming and effortful to support MC/SRL development along with delivering mathematics content and their implementation could fluctuate. It is not possible to completely account for the lower effects of dissertations in this review, partly because they all originated in the United States and used a range of assessments to check mathematics outcomes. Issues related to geographic context or assessment comparability could also play a role and should be considered by users of the review.

Next, it was important to see if there would be age-related distinctions in effects similar to those found in previous reviews. For example, Dignath & Büttner (2008) found MC/SRL-training effects on mathematics much higher in primary school pupils (ES=0.96) than in secondary pupils (ES=0.23). As discussed above, large-scale studies with standardised assessments tend to find higher effects in mathematics in younger students, but they are relatively stable after about 6th grade (Kraft, 2020, p. 249). On the other hand, some theories emphasise that cognitive maturity and subject domain knowledge are necessary for effective MC/SRL functioning, and both should be more developed in older students. In their review of

ICT-based mathematics programmes with MC support, Verschaffel et al. (2019), suggested some MC training might overtax the working memory of younger children or pupils with special needs (p. 5). Therefore, I did not have a strong expectation about the relationship between age and MC/SRL-programme effects in this review. In the current review, most studies focused on upper-primary and middle-school pupils. Based on the moderator analysis, there was no relationship demonstrated between age and effect size. There could be several reasons for this. As demonstrated in the narrative synthesis, even within the same age and grade ranges there were a variety of approaches to training MC/SRL skills. Studies also varied with regard to the dose or length of the intervention, discussed below. These variations might obscure real variation in effect due to age. It is also possible that ability-based variation could also obscure the effects related to age. Some studies reported that the intervention group included at-risk students, those with special educational needs, or gifted or high-achieving students. Other reports specified that participants were average-achieving or of mixed ability, and some studies did not report abilities or needs designations for students. Thus, it could be difficult to disentangle age- and ability-related factors, especially if assessments were also tailored to the participants. In addition, included studies often did not discuss whether they optimised task difficulty to encourage practice of the new MC/SRL skills. If tasks are too easy, pupils may not see a benefit, and if too hard they may simply give up (Efklides, Samara, & Petropoulou, 1999; Mevarech & Kramarski, 1997; Rellinger et al., 1995).

It could also be the case that the data extraction methods did not capture real differences due to age. I coded the average age based on the range of ages or the grade-level reported. In a few cases, the study included multiple grade levels, and these ages were averaged together if all participants took the same assessment, and they were entered into a fixed-effect meta-analysis to get the study-level effect. For the moderator analysis, only one average age from each study was entered, thus it only considers variation by age between studies but not within studies. This reflects the choice to use only one effect size from each study in the meta-analysis, but it could mask meaningful differences in effects by age. Because the review includes studies from international contexts, it was not always clear which ages would be included in the year groups reported, and my conversions could be inaccurate in some cases. To determine whether there is a real difference in effect due to age, it would be helpful to compare studies using the same interventions and assessments, with each study focusing on a different year group of participants. Only if other elements were held constant and reported clearly, would a future meta-analysis yield robust results on this question.

The final potential moderator of effect that was considered was the “dose” of the MC/SRL intervention, operationalised as weeks of training for students. Because a flexible and strategic MC/SRL mindset can take time for students to develop, and because impacts on classroom climate and productive dialogue might also be important for intervention effectiveness, I predicted that studies with a longer “dose” of MC/SRL training would produce higher effects. However, the moderator analyses did not demonstrate this relationship, for which there may be several reasons. First, it might be possible that there is no relationship between MC/SRL training dose and its impact on mathematics performance. That is, a longer intervention might have no greater impact than a shorter one. This could be the case if MC/SRL concepts and behaviours operate as a “threshold” (Meyer & Land, 2003), where students simply need to have minimal exposure to them in order to be prompted to be more strategic in their learning and performance. This possibility cannot be confirmed or disproved by the current review, but it could be explored in further studies. In no studies from this review did the authors compare the effects of longer vs. shorter interventions, so this is an area open for future research. On the other hand, the lack of a clear correlation between dose in weeks and effect size does not necessarily mean there is no relationship, but it could mean that other factors, such as the specific activities and strategies of the intervention, had a more definite connection to the outcome of interest. Moderators could interact with one another, so intervention dose might make more difference to outcomes for certain ages or types of interventions (e.g., structured problem-solving vs. general metacognition).

A third possibility is that methodological variations within the primary studies obscure a direct relationship between dose and effect. For example, there is reason to believe that the assessment tools used were already scaled to the length of the intervention, based on the mathematics content covered. Within the review, assessments ranged from a few items testing proficiency with a single mathematical concept or problem-type, to district- or state-level unit tests, to large-scale, standardised assessments, such as Measures of Academic Progress (MAP), which test a range of mathematics subskills. The choice of assessment seems to align in many cases with content covered during the intervention period, so that greater learning growth in a broad scope intervention might equate to the same effect size as a shorter intervention with a narrower assessment as an endpoint. If future research aimed to clarify this, it would be preferable to compare interventions of different lengths all being tested with the same assessment. Large-scale standardised and adaptive tests would be ideal for this, and they would also enable the researchers to avoid potential bias in tailoring an assessment to the intervention under study (i.e., treatment-inherency, Slavin & Madden, 2011; Rogde et al., 2021).

To determine the relative contributions to effect of different factors of MC/SRL programmes, future meta-analyses could include a meta-regression, as was done in Dignath and Büttner (2008). I decided against employing this approach for the current review, given the relatively low correlations (psuedo R^2) found for each moderator analysed. It is likely that other factors not examined as moderators could also have an important influence on effects, such as the nature of the comparison group(s) and the specific MC/SRL strategies trained. Previous reviews frequently categorised interventions based on whether they trained cognitive, metacognitive, or motivational strategies, but I found that it was a challenge to distinguish cognitive and metacognitive strategies–for example, how should we classify asking pupils to draw a picture representing a mathematical task?--and motivational strategies were frequently combined with the other two. High variation in how MC/SRL programmes were implemented and reported obfuscates the analysis of moderators, which is considered further in the next section on study limitations.

5.3 Narrative synthesis findings

This review demonstrates that it is possible to implement effective MC/SRL programmes in a variety of contexts with different resources and constraints, and to meet the needs of specific learning communities. The geographic spread of included studies was high, with the highest numbers from the US, the Middle East, and Europe, and fewer from East Asia and Oceania. There are potential gaps in coverage, with no included reports from Africa, South America, and the Indian subcontinent. Requiring reports to be in English may have affected this, but several English-majority nations were also not represented, namely Canada, Australia, and New Zealand. The geographic spread of the current review differed from previous reviews (e.g., Higgins et al., 2005), and the fact that all dissertations/theses were from the US complicates analysis of geographical spread. It is clear the worldwide interest in MC/SRL programmes has not waned since previous reviews, but more work is needed to understand the nature of “traditional” mathematics teaching in each context and how MC/SRL implementation could vary.

Various school types were also featured, such as public, private, and charter schools, as well as students of different abilities, ethnicities, genders and language and SES backgrounds. US-based research tended to highlight these characteristics more, and they could be more salient there given the diversity of schooling contexts. Even when pupil demographics were reported, the expected relationship with outcomes was sometimes opaque in study reports, except that a few studies focused specifically on advanced or gifted learners (e.g., Rizk, Attia, & Al-Jundi, 2017), those with special needs (e.g., Kang, 2010), or second language speakers (e.g.

Morales, 2016). Several included studies reported outcomes by gender or featured gender-segregated classes. Gender-based differences in outcomes have been seen, including in several studies in this review (e.g., Edwards, 2008; Falco, 2008; Jackson Jackson, 2012). In particular, there has been a concern that females have lower achievement in mathematics by the end of compulsory schooling, despite starting school on an even footing with males, and that gender may impact how students approach mathematics tasks (Seegers & Boekaerts, 1996). The reasons for these discrepancies may be complex and involve both social and individual factors (Entwistle, Alexander, & Olsen, 1994; Harris et al., 2021), but MC/SRL training that encourages all students to explore and take ownership of their beliefs about mathematics could be important. Future MC/SRL-based studies could explore students' mathematical commitments and self-efficacy vis a vis gender or other identity categories, with an explicit focus on developing meta-affect, self-efficacy, and healthy attributions through hands-on mathematical production. As Schoenfeld (2020, p. 1173) states: "Our classrooms must be environments in which all students are empowered to engage meaningfully in mathematical practices, for such engagement is the source of agency and identity."

The MC/SRL-related activities and strategies were coded in detail based on included reports, and these varied widely. Most studies reported a multi-stage MC/SRL focus, with pre-, during-, and post-task strategies, such as setting goals, planning a task approach, monitoring performance, and assessing task completion or responding to correctness feedback. However, many strategies operated at the task or problem-level with only a few studies examining strategies for domain-level learning. Most interventions focused on structured problem-solving, discussion of task approaches or understanding, and calibration-type exercises, with many interventions including graphing, written reflection, or behavioural and affective regulation, and only a few interventions featuring study strategies, play-based activities, or free mathematical exploration. More commonly, the MC/SRL programmes were "infused" (Higgins et al., 2004) within specific units of curriculum rather than being used as a supplemental or "enrichment" activity.[39] The simplicity of the required materials and activities–most interventions used worksheets–and their implementation by the regular classroom teachers indicate that MC/SRL programmes can be done at a small scale, for a low cost, and with minimal resources and disruption to normal classroom activities. It is less apparent which versions of MC/SRL training are more practical, acceptable, or effective in specific contexts. Some interventions were more comprehensive than others, and in some cases the intervention activities were not described in

[39] As a reminder, interventions needed to be done during normal school hours to be included in the review so that standard practice would be an appropriate comparison.

detail. This leads into the potential limitations of the included studies and the systematic review itself.

5.4 Limitations of the research

No single study, even a well-planned and executed systematic review, can be the last word on an area of research or practice. In addition to this general limitation, this section covers specific limitations of the general research approach, the implementation of the methods, and the included studies that could muddle the interpretation or application of the findings. A major issue was underreporting of the primary studies, specifically related to the intervention components, but the mathematics assessments and quality-control measures could also be improved. Although I sought to implement the systematic review methods in line with accepted guidance, there are undoubtedly ways this could have fallen short. Finally, I reflect on conceptual and design issues for the systematic review, which it would be profitable to revise should a future review be planned.

5.4.1 Limitations of included studies

First, the included studies contributed incomplete information regarding some categories of interest for this review. All included studies met the pre-planned criteria which I devised to ensure they could be used in the meta-analysis, such as those relating to the design and assessments used as well as naming metacognition or self-regulated learning as part of the theoretical basis for the interventions. In extracting the statistical information to generate the effect sizes, I also found great variations in how studies reported outcomes, with around a third of studies not reporting means and standard deviations clearly for MC/SRL vs. Non-MC/SRL groups. I was able to use alternate effect size calculations to retain such studies. Arguably, a more critical omission was when studies failed to clearly and completely report the intervention activities, as recommended by Hoffmann et al. (2014). A research report can only be properly interpreted if readers understand precisely what was tried, and education research, ostensibly intended to improve practice, should take extra care in this. Within the current review, some interventions were simpler, such as writing reflective statements for the last few minutes of a lesson (e.g., O'Neal, 2013), while others were complex and included several MC/SRL components (e.g., Byrd, 2019; Falco, 2008). Only some studies included enough detail in the main report or a supplement to enable replication of the approach. Yet even where these were given, there was often little description about how students and teachers worked with and responded to the MC/SRL activities. Where the implementation was done by external personnel or using technology, regular teachers' involvement in the MC/SRL interventions was mostly unclear. As most interventions were not delivered during every class period, it also would have been helpful to know to what extent the MC/SRL practices permeated into “normal” teaching, and more detail about the teaching in control groups would have illuminated the nature of the comparisons being made in the study. MC/SRL programmes may operate at both an individual and social level within a classroom, but most of the reported outcomes from studies in this review relate to the former while only occasionally touching on the latter. Based on reporting, it seems that the MC/SRL “feedback loops” (Dent & Koenka, 2016, p. 427) are presumed rather than confirmed, yet these are precisely the types of processes that could prove challenging to support in a new context. At a minimum, teachers and pupils should be surveyed about their perspectives of the intervention at several time points. It would also be ideal to observe class sessions prior to, during, and after the intervention period. Both of these measures would enable a more thorough representation of the potential social mechanisms of effect in MC/SRL interventions and show growth over time. This could be invaluable information for replication or application of the studies.

Another limitation of the included studies relates to the assessment tools used to report mathematics outcomes, which were combined in the meta-analysis and are shown in Appendix 7. The objective was to see what difference MC/SRL-training would make on more traditional mathematics assessments, and when multiple assessments were used, I chose those that were longer and/or of broader scope, and ideally externally developed and validated with a sample other than the participants of the study. Not all studies included such an assessment, so as a minimum I required studies to include at least one assessment where the focus was primarily on correct calculations, even though this sometimes represented a narrower mathematics-achievement construct. Still, even with these efforts, there could be important ways that assessment differed. For example, many assessments focused on word problems, which require interpreting mathematical language, determining which calculations to perform, and executing the calculations correctly. In other assessments, questions were posed numerically, with less need for test-takers to determine which operations to perform. In addition to potential lack of comparability, as mentioned in the results chapter, it was a challenge to glean from reports how difficult the assessments were for the pupils in the study, which could have impacted results in at least two ways. If the problems were too difficult or too easy, they might not trigger the application of the trained MC/SRL-skills, and the test might show either a “floor” or “ceiling” effect, meaning they would have less discrimination power and the use of the standard deviation as the effect size unit might be less appropriate. In addition, some studies only reported posttest scores adjusted for baseline, but this would again be less appropriate if the assessment difficulty were not calibrated well. In fact, the same assessment could be too difficult at pretest and too easy at posttest, and not able to show meaningful differences between individuals or between intervention and control groups. The last assessment-related point is that outcomes may have been scaled to the mathematics content under study during the intervention period. For example, effects of a semester-long intervention could have been assessed with a final exam or standardised achievement test while those from a shorter intervention could be assessed with a much shorter test of narrower scope, including, perhaps only a few word problems. This possibility became apparent following the conclusion of the data collection and the execution of the systematic review, especially upon finding no effect size difference based on the dose of the interventions. In future reviews, it would be productive to systematically investigate this. Additionally, researchers should strongly consider discontinuing the use of purpose-built assessments or those adapted for the research and instead using existing, larger scope assessments. This is discussed further below in the recommendations section.

The final limitation related to the primary studies of this review is the lack of attention to quality-related factors, with only around a third of studies reporting adherence to each of the eleven criteria I checked. CONSORT (Grant et al., 2018) and other guidelines urge researchers to report key aspects of study design and implementation so users can determine how robustly they addressed the research questions. In addition, there is a rich methods literature regarding how research should be planned in the first place. Conducting research in a complex and constantly shifting social setting, like a school or classroom, means that plans may be altered, but such changes should still be reported. For this systematic review, I intended to use a holistic quality assessment similar to Gorard's "sieve" (2014, p. 54) to rate studies. This was intended to assist research users in understanding the strength of the evidence about the combined effect size, as well as selecting high-quality primary studies from the review for detailed follow-up. In their EPPI-Centre reviews on "thinking skills," Higgins et al. (2004, 2005) use their own rating system and label most included studies as medium "weight of evidence" regarding the appropriateness of designs and methods to answer the research questions. However, they also found many reports lacking such details (Higgins et al., 2005, p. 23). Similar issues appeared in the data-extraction for the current review, with numerous reports omitting details on how participants were selected and sample maintenance issues, like attrition, cross-over, and loss to follow-up. It was often unclear how teachers were recruited and what their roles were if they were not the main leaders of the intervention. Issues like participants' knowledge of condition and blinded/masked outcome assessment were rarely mentioned, nor was it discussed whether participants remained in the study voluntarily or were apprehensive about the consequences of participating or not participating. The latter are not only ethics issues because they may affect the causality logic of the study. If other factors cannot be ruled out, it is more difficult to conclude that the MC/SRL training led to the effects. For example, schools could have selected their best teachers as intervention leaders, and teachers could have invested much time and effort into making the programmes effective. Pupils could also try harder if they know they are in a special programme. Although many studies reported using random assignment and sometimes random sampling, it was often not stated how this was done or by whom. Even when classes or schools were randomly assigned to study groups, with a low number of assigned units there could still be baseline imbalances, and researchers frequently analysed outcomes at the individual pupil level without correcting for nested or correlated data. When reporting on quality factors is insufficient, Gorard (2014, p. 49) states research may be seen as untrustworthy. Thus, there is an imperative for studies, even those not strictly following an RCT approach, to report how quality-control factors were designed and if changes were made during the study. Such reporting should be considered a strength not a weakness, but with the omissions noted, the results of this systematic review are less conclusive. Gorard (2014, p. 48) warns that, without accounting for quality, there is a risk that ". . . weak evidence will be bundled along with strong evidence [in syntheses], leading to invalid and possibly dangerously misleading conclusions."

5.4.2 Limitations of the synthesis methods

In design and executing the research methods, I sought to follow key guidelines, such as PRISMA, the Campbell Collaboration handbook, and chapters by experienced reviewers (e.g., Torgerson, Hall, and Light, 2012). These resources helped specify the inclusion and exclusion criteria, search strings, and screening process to make the search as exhaustive and replicable as possible, while still being practical for me to execute as an independent researcher. I used an iterative and documented process to refine the searches and databases used. I sought expert advice for each stage of the review, and I incorporated double-rating at two points to feed back on my methods. For the meta-analysis, I was able to employ free, user-friendly tools, such as the Campbell Collaboration effect size calculators and Meta-Essentials, demonstrating that researchers with various resources and training backgrounds can contribute to synthesis research. Overall, I was able to address the main research question regarding the effects of MC/SRL programmes on mathematics outcomes, based on a review of 60 primary studies. I was also able to explore potential moderators, even though the results were not as clear as the combined effect.

Nevertheless, there are methodological limitations within the current review. In any research endeavour, it could be useful to have a team of researchers pooling their resources and strengths. Several methods texts for systematic reviews advise using such a team (e.g., Peterson et al., 2017; Torgerson, Hall, & Light, 2012), yet I designed and executed the search and screening, data extraction, and meta-analysis working mainly independently. Even with careful planning and expert guidance, there could be some areas in which “best practices” were not followed. As I was learning the synthesis process, I could not anticipate how long each stage would take, and it is possible that some stages, such as the screening of items for inclusion, took impractically long. One reason may be that I was not aware of how much my search outlets overlapped in their coverage, and so there were more duplicates to screen out, which sometimes needed to be done manually. I also sought to account for the reasons items were excluded, even though many reviews do not do this exhaustively, and this undoubtedly added to the screening stage. In doing so, I also became aware that design issues were a major reason items needed to be excluded. Although I had included several terms to specify causal designs in my search, it still did not discriminate well between observational studies and those with active manipulation, for example. If the search had been refined to include only experimental and quasi-experimental studies, I would have spent less time screening out items and had more time for comprehensive qualitative analysis. Partly due to practicality reasons, I also excluded studies that could not contribute to the meta-analysis because they reported outcomes inappropriately, but I could have still included these in the narrative synthesis or in a broader, “scoping review” of the research area (Peterson et al., 2017). Doing so would mean the findings about the nature of MC/SRL-training programmes would have been more comprehensive, but it might have been less clear how such findings would link to effectiveness. Similarly, I made the choice not to meta-analyse outcomes on MC/SRL-related measures, even though many studies included them. This was done to focus on the mathematics outcomes more directly, but some researchers might consider this a limitation since the mechanisms of effectiveness are potentially obscured. Finally, as mentioned above, included studies varied widely in their quality, and this may have contributed to the overall heterogeneity in the combined effect and complicated the moderator analysis.

5.4.3 Limitations of the research topic

For this review, I sought to combine effects from a broad range of MC/SRL programmes. I combined studies with an explicit metacognitive theoretical basis with those based on self-regulated learning because I believed these both operated in similar ways within a constructive teaching and learning framework. While there has been disagreement regarding which construct operates at a higher level subsuming the other (Verschaffel, Depaepe, & Mevarech, 2019, p. 2), both theories of metacognition (e.g. Brown, 1977; Flavell, 1979; Winne, 1996; Pintrich, 2002) and self-regulated learning (e.g. Zimmerman, 1990; Bandura, 1991; Boekaerts, 1999) posit that learners should develop greater awareness of themselves, their beliefs and strategies, and how these fit the task or domain, and they should incorporate different types of feedback to optimise how they function within a task or domain. There are nuanced distinctions between these theories. For example, metacognition might not always be considered as including motivational or affective components[40]. In practice, however, there seems to be little difference in the types of activities that could be included in metacognitive vs. self-regulated learning interventions, and many of the included studies for this review mentioned both theories. There were certainly differences in the activities the researchers chose, but these differences were not consistently related to specific MC vs. SRL theories. The meta-analysis showed that MC/SRL interventions of various forms could be effective on mathematics outcomes, yet there was substantial heterogeneity reflected in the combined effect. Differences in foundational theories or the intervention activities and MC/SRL strategies could be linked to differences in effects, but this could not be determined due to the ways studies implemented and reported them. Because some approaches combine multiple MC/SRL approaches in a comprehensive programme, it could be profitable to limit a future review to only such studies and exclude those that implemented only one or two strategies in a limited way or that focused on specific sub-constructs, such as self-reflection, attributions, calibration, motivation, or behavioural control. Even though these may be key components of a comprehensive MC/SRL approach, by themselves they may fail to catalyse the positive feedback loop referred to above, and this could lead to confusion among practitioners. This idea is returned to below in the section on recommendations based on the research.

[40] Yet it is worth considering Panadero's (2017) discussion of Efklides's MASRL model, which he says bridges this gap. Noticing one's affective or motivational state could also fit Flavell's category of metacognitive experiences (1979, p. 908)

5.5 Recommendations from the research

As education research, this systematic synthesis has aimed to contribute to improved educational practices, thus I close with recommendations for teachers and school-leaders, as well as for the researchers and theorists charting new paths in the MC/SRL landscape. First, academics should design, carry-out, and report MC/SRL research with practice in mind. For example, it may be less critical to highlight distinctions between metacognition and self-regulated learning than it is to show their symbiotic relationships, and particularly how they are used in teaching in both rich and concrete ways. Researchers should strive to flesh out holistic MC/SRL practices, while showing how these can make a difference to traditional, broad scope assessments. On a positive note, this review strongly demonstrates the potential for good outcomes when MC/SRL programmes are implemented in mathematics learning, but educational stakeholders need to be prepared to make deep and lasting changes in the classroom. Each of these recommendations is explained further below.

5.5.1 Recommendations for research and theory

As discussed at length in the results chapter and above, understanding of the core components of the interventions and designs of the review studies was sometimes hampered by reporting issues. Although several guidelines state that such reporting is essential (e.g., Hoffmann et al., 2014; APA, 2008), these have not been applied consistently within education research. In many cases, reports are quiet on sampling and allocation to groups and whether programmes were faithfully implemented, aspects of research that are admittedly arduous to control in a school context. Some reports featured only a short paragraph on the intervention itself, while theoretical discussions and quantitative analyses proceeded for several pages (e.g., Abdullah, Hali, & Zakaria, 2014) or the discussions of the intervention activities are too vague to be repeatable (e.g., Aminah et al., 2018). On the other hand, some reporters included high detail about the intervention in the main report (e.g., Lee, Yeo, & Hong, 2014) or a supplement or appendix (e.g., Wang et al.), such that practitioners would be able to implement them without additional resources or training. My recommendation is that researchers follow these examples and report MC/SRL interventions as fully as possible and reduce other report sections if necessary to facilitate this. MC/SRL theories, for example, have been discussed at length and may need only a short summary within the research rationale, and in well-designed studies the outcomes may be reported with only a few carefully chosen analyses (Gorard, 2013, p. 203). Researchers can also link to additional files hosted online or take advantage of electronic supplement options from publishers if the intervention is too detailed. Logically, the length of the intervention description should also be matched to the complexity of the intervention. A simple intervention, such as post-lesson reflection, may be described in a few sentences, while a complex multistage intervention may take several pages. Samples of activities and dialogue from the in-use intervention could be valuable in helping practitioners operationalise key MC/SRL concepts. The latter type of reporting, along with samples of student work, can also confirm expected mental and social processes, and researchers should discuss whether they expected the MC/SRL-training to carry-over to normal teaching sessions and if this happened. All of these reporting recommendations would help to draw a much richer picture of the co-construction of metacognitive and mathematical knowledge, beliefs, and dispositions as it occurs in the context of MC/SRL programmes in schools.

Based on the current review, another recommendation is to strengthen MC/SRL-based studies to engender more robust outcomes assessment. In some cases, for example, researchers employed multiple intervention groups compared against the same control group. Following the lead of Dignath and Büttner (2008, p. 241), I reduced the size of control groups and combined all MC/SRL groups in the effect size calculation. However, the use of multiple variations of MC/SRL training in addition to a BAU control indicates a lack of consideration regarding the “gap” in current knowledge and possibly a lack of “equipoise” to justify undertaking the study (Lilford & Jackson, 1995; Gorard, 2013, p. 134). Ethically, studies should always try the best versions of their interventions, since real student outcomes are at stake, and eliminating extraneous groups would simplify the quantitative analysis. The current review did not include studies without a non-MC/SRL group, but based on my findings and those from previous reviews, it is perhaps more appropriate to eliminate the BAU control group, especially if the teaching approach is very traditional and focused on guided practice and error avoidance. At the same time, there is value in real replication research, though my review does not show any examples of such. Each time the IMPROVE (e.g., Kramarski & Gutman, 2006) or “Red Light, Purple Light” programmes were used, for example, there were changes, and this makes comparisons a challenge. Also problematic is the use of researcher-developed or adapted assessment tools, which could be overly sensitive to the intervention teaching or mask differences in effects from different intervention lengths, as discussed above. I strongly recommend that MC/SRL programmes utilise and report standardised or externally-validated broad-scope assessments to check academic outcomes, possibly in addition to more proximal, treatment-responsive assessments. Broad-scope assessments are the ones stakeholders, such as parents, administrators, and policy-makers, are most interested in, and seeking to show an impact on such assessments could also create positive washback by encouraging more comprehensive and extended MC/SRL programmes rather than training isolated strategies.

Regarding theoretical recommendations, the lack of alignment between stated MC/SRL theories and specific activities and strategies is concerning, as discussed above. According to Zimmerman (1986, p. 312), ". . . theories are useful heuristically to the degree that they raise specific issues that can be resolved through research." In this research area, theorists might offer the most assistance by clarifying what key elements should comprise MC/SRL training. I would propose this should be comprehensive and reflect the nature of well-known SRL models, as presented in Panadero (2017). Each model clearly shows a cycle or feedback loop involving regulation sites (e.g., person, environment, behaviour, Zimmerman, 1989, cited in Panadero, 2017, p. 3) task phases (forethought, performance, self-reflection, Zimmerman & Moylan, 2009, cited in Panadero, 2017, p. 5), or cognitive, metacognitive, and motivational elements (Boekaerts, 1996, cited in Panadero, 2017, p. 7). As argued above, individual metacognitive or regulation strategies could influence such cycles, but incorporating multiple elements from different parts of the cycle will more likely lead to genuine and sustainable MC/SRL development.

5.5.2 Recommendations for practice

This chapter concludes by presenting recommendations for improving practice, which should be the ultimate goal of education research. The results of the systematic review and meta-analysis offer encouragement to teachers and school-leaders interested in MC/SRL approaches. They can be implemented effectively in many ways, and they can be adapted to the various needs of mathematics classrooms and learners. Reviewing MC/SRL theories and preparing to implement an intervention based on them could assist teachers in developing their own reflective practices and informal theories of students' thinking, which are needed to plan instruction and assessment. Along with this, teachers' self-efficacy as skilled professionals should increase if they are given scope to make strategic decisions in the classroom just as students are (Baumfield et al., 2009). On the other hand, although MC/SRL programmes can be low-cost, teachers and administrators need to be prepared to make deep shifts in classroom roles and culture and to persist in MC/SRL activities for a sufficient period to see improved outcomes, with all the other demands placed on them. Task-related strategies like structured problem-solving with metacognitive prompts (e.g., IMPROVE) could lead to faster gains in achievement, but more comprehensive programmes focused on mathematics identity, motivations, and affect may be needed to change expectations about who can be successful in mathematics and how. With such approaches, schools can support mathematics learning as well as the non-academic knowledge and skills to “make wise and thoughtful life decisions” (Flavell, 1979, p. 910).

Appendix 1. Central concepts, MC/SRL definitions, and search terms of previous MC/SRL reviews

Review	Core construct(s)	Definition or description of MC/SRL, with illustrative quotations	Search terms
Hattie, Biggs, & Purdie (1996)	Study skills	"These interventions have aimed at enhancing motivation, mnemonic skills, self-regulation, study-related skills such as time management, and even general ability itself; creating positive attitudes toward both content and context; and minimizing learning pathologies" (pp. 99-100).	study skills, learning strategies, learning processes, cognitive style, study habits, cognitive strategies, cognitive processes, learning style, metacognitive skills, thinking skills
Higgins et al. (2004, 2005)	Thinking skills	". . .approaches or programmes which identify for learners translatable mental processes and/or which require learners to plan, describe and evaluate their thinking and learning. These can therefore be characterised as approaches or programmes which: • require learners to articulate and evaluate specific learning approaches • identify specific cognitive, affective or conative processes that are amenable to instruction" (pp. 7-8).	thinking, thinking skills, thinking skills program(me), thinking strategies, critical thinking, critical thinking skills, creative thinking skills, higher order thinking skills (HOTS), metacognition, metacognitive, meta-cognitive/ition community of inquiry/enquiry/learners, transfer, near-transfer, far-transfer, bridging, teaching for transfer, reasoning, argument, Socratic questioning, mediated learning, Instrumental Enrichment/ Feuerstein, Somerset Thinking Skills / Blagg, Top Ten Thinking Tactics / Lake, Cognitive Acceleration in Science/Maths/Technology Education (CASE/CAME/CATE) / Adey, Shayer, Adhami, Philosophy for/with Children (P4C) / Lipman, Thinking Actively in a Social Context (TASC) / Wallace, Activating Children's Thinking Skills (ACTS) / McGuinness, CoRT (Cognitive Research Trust), Six Thinking Hats / deBono, Storywise, Philosophy with Picture Books / Murris, Reason!Able / van Gelder

Review	Core construct(s)	Definition or description of MC/SRL, with illustrative quotations	Search terms
Dignath, Büttner, & Langfeldt, (2008); Dignath, & Büttner, (2008)	Self-regulated learning	“. . . self-regulated learning is characterized as an interaction of cognitive, metacognitive and motivational processes, which work together during information processing. . . . the first level consists of cognitive strategies, which refer directly to information processing. The second level relates to the use of metacognitive strategies aiming at the regulation of the learning process. The third level illustrates the maintenance of motivation, which is characterized by the willingness of independent goal setting, self-activation, as well as adaptive coping with success and failure . . .” (Dignath, Büttner, & Langfeldt, 2008, p. 104).	study skills, learning strategies, self-regulatory strategies, self-regulatory skills, metacognition, metacognitive skills, metacognitive strategies, self-regulated learning, motivational skills, self-motivation, life long learning, learning to learn, thinking skills, learning processes, cognitive style, cognitive strategies, study habits, learning style, cognitive processes, goal-directed behaviour, self-monitoring, goal-setting, self-control, self-determination, self-management, organizational skills
Donker et al. (2014); de Boer, Donker, & van der Werf, (2014)	Self-regulated learning, metacognition	“Self-regulated learners are students who are capable of supporting their own learning processes by applying domain appropriate learning strategies. . . . Self-regulated learning can be described as: ‘‘an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate and control their cognition, motivation and behavior, guided and constrained by their goals and the contextual features in the environment’’ (Pintrich, 2000, p. 453). In short: students who are able to self-regulate their learning are active, responsible learners who act purposefully (i.e., use learning strategies) to achieve their academic goals. To this end, they need metacognitive knowledge; knowledge and awareness about their own cognition . . .” (p. 2)	metacognit*, self-reg*

Review	Core construct(s)	Definition or description of MC/SRL, with illustrative quotations	Search terms
Dent & Koenka, (2016)	Self-regulated learning, metacognitive processes	"Specifically, *learning is self-regulated to the extent that students are motivationally, cognitively, and behaviorally engaged in the academic task* (Zimmerman 1986)" (p. 426). ". . . a working definition that incorporates many of these perspectives was proposed by Pintrich (2000), who argued that self-regulated learning is an active, constructive process whereby students set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features of their environment (p. 453)" (p. 427). " . . . Because metacognitive processes explain how self-regulation occurs, they are a source of both consistency and controversy among researchers from different theoretical perspectives. However, a consensus has begun to emerge around five of these metacognitive processes: goal setting, planning, self-monitoring, self-control, and self-evaluation" (p. 428).	self regulated learning, metacognitive strategies, metacognitive skills, metacognition, cognitive strategies, learning strategies, academic, GPA, class grade, achievement test, academic achievement, academic performance, course grade, standardized test, achievement test, LASSI, MSLQ, PALS, ILP, ASI, ASSI, SRLIS, Mandinach 1984, Winne 1982

Review	Core construct(s)	Definition or description of MC/SRL, with illustrative quotations	Search terms
Ergen & Kanadli (2017)	Self-regulated learning, metacognition	"Self-regulated learning is defined as an active and constructive process in which individuals set their own learning goals, regulate their cognition, motivation, and behaviours, and are directed and limited by their own goals and contextual features around (Pintrich 2000) . . . Students getting to know themselves . . . [is] a process that is associated with metacognitive skills, acquiring knowledge with cognitive skills, and obtaining the ability to motivated [sic] themselves and manage their environment effectively. For this reason, self-regulated learning model is explained in four categories: cognitive, metacognitive, resource management, and motivational strategies (Pintrich & De Groot, 1990; Pintrich, 1999)" (p. 56).	(Turkish and English) self-regulating learning, self regulated learning, learning strategies, metacognitive strategies, metacognition, social cognitive theory, academic success, academic achievement

Review	Core construct(s)	Definition or description of MC/SRL, with illustrative quotations	Search terms
de Boer et al. (2018)	Self-regulated learning, metacognitive strategy instruction	"Pintrich (2000) described self-regulated learning as 'an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment' (p. 453). A typical characteristic of self-regulated learners is that they use strategies which facilitate and enhance their learning process and consequently their academic performance (Zimmerman, 1986; 2002)." (p. 98) "Metacognitive strategies are methods which facilitate and regulate cognition. Because the use of these strategies involves monitoring and controlling one's own learning, including the application of cognitive strategies, metacognitive strategies are considered higher-order skills, and are more difficult to teach than cognitive strategies . . . Finally, management strategies are applied to deal with the context of the learning environment" (p. 99).	self-reg*, metacognit*, learning strat*, study strat*, learning skill*, study skill*, strat* use, strat* instruction, follow-up, delayed, maintenance, retention, long-term, intervention, program, treatment, instruction, experiment, training

Review	Core construct(s)	Definition or description of MC/SRL, with illustrative quotations	Search terms
Lee et al (2018)	Metacognitive training, algebraic reasoning	"Metacognition includes students' thought processes and beliefs that enable them to regulate their learning activities (Schoenfeld, 1987). . . .General metacognition refers to being able to regulate problem-solving processes regardless of the specific situation. Domain-specific metacognition focuses on the unique characteristics of each situation among diverse/complicated situations (Kramarski & Mevarech, 2003). During the metacognitive processes, both knowledge and cognitive skills are planned, monitored, analyzed, evaluated, and reflected by students based on their own goals" (p. 43)	metacognitive training, metacognitive guidance, metacognitive instruction, metacognition, algebraic reasoning, algebraic thinking, algebraic achievement
Perry, Lundie, & Golder (2019)	Metacognition	"the Organisation for Economic Co-operation and Development (OECD), [states] that metacognition is,…a second or higher-order thinking process which involves active control over cognitive processes. . . .The majority of researchers separate metacognitive knowledge from metacognitive skills. . . In addition to this, we accept the three level model suggested by Donker et al. who recognise 'an interaction of cognitive, metacognitive and motivational processes, which work together during information processing' (Donker et al. 2014)" (p. 485).	*Search terms not reported.*

Review	Core construct(s)	Definition or description of MC/SRL, with illustrative quotations	Search terms
Verschaffel, Depaepe, & Mevarech (2019)	Metacognition, ICT	"Examples of metacognitive components are planning, monitoring, control, and reflection . . . From the very beginning, researchers . . . distinguished between two closely interrelated components of metacognition: (1) knowledge of cognition (e.g., knowledge about the task, strategies appropriate for solving the task, and personal characteristics relevant to the task) and (2) regulation of cognition (e.g., monitoring, control, and reflection). (The term 'metacognition' is closely related to the term 'self-regulation.' . . . we take the former perspective, implying that we conceive metacognition as the most general concept and self-regulation as the second, regulatory component of metacognition.)" (pp. 1-2).	metacognition, self-regulation, mathematics, online learning, computer-based learning
Wang & Sperling (2020)	Self-regulated learning	"Self-regulated learners are active agents who use a repertoire of knowledge and strategies to regulate their learning adaptively and efficiently (Zimmerman, 1990, 2002; Schraw and Moshman, 1995). Self-regulated learners also examine their strengths and weakness against academic task standards in order to set appropriate goals, deploy strategies, adapt to varying environments, and to overcome obstacles. . ." (p. 1).	self-regulation, metacognition, learning strategies, mathematics, mathematics skills, school-based intervention, intervention

Appendix 2. Protocol for the search and screening based on PRISMA-P

1a Title: Effects of metacognitive and self-regulatory interventions on mathematics achievement in primary and secondary students: Protocol for a systematic review and meta-analysis

1b Update of previous review?: No

2 Registration of protocol: None

3a/3b Authors

First/Corresponding author and guarantor: Loraine Hitt (Durham University Ed.D. candidate) Email: l.e.hitt@durham.ac.uk; Mailing address: 105 Wofford Lane, Conway, SC, USA 29526

Supervisor: Professor Steven Higgins, Durham University School of Education, Email: s.e.higgins@durham.ac.uk

4 Amendments

In the event changes to this protocol become necessary, they will be listed in this section along with the date and rationale. The supervisor will review any proposed changes for appropriateness.

5a Support

The first author has been provided with a free EPPI-Reviewer account by the supervisor. EPPI-Reviewer is a multi-function tool specifically designed for collecting, screening, extracting data, and analyzing data for systematic reviews. The first author will otherwise be self-supported in this review.

5b/5c Sponsorship

The School of Education at the University of Durham, UK, will advise and oversee the research. The first author assumes responsibility for the methods, ethical conduct, and intellectual contributions of the review, along with any potential oversights, omissions, or errors.

6 Rationale

Interventions based on metacognition and self-regulation have shown potential to raise general academic achievement for lower costs than other types of interventions. Within the domain of mathematics, it is predicted that metacognitive interventions may offer an additional benefit beyond teaching strategies for mathematical problem-solving per se. This review is aimed at providing teachers of primary and secondary mathematics information about the types of metacognitive interventions likely to be good investments of classroom time and other resources. To do this, the review seeks to locate and summarise the most trustworthy evidence using clear and replicable techniques to control for potential biases.

7 Objectives

This review will seek to address the following research questions.

1. During the last 15 years, what has been the effect of interventions based on theories of metacognition or self-regulation on the mathematics achievement/proficiency of school-aged learners?
2. What specific factors, if any, are correlated with higher effectiveness for such interventions?

8 Eligibility criteria (See table below for full criteria)

a. **Participants:** Given that this review aims to support educational practice, the focus of this review is on intervention studies carried out in regular school settings where mathematics is a prescribed part of the curriculum, regardless of who carried out the intervention with students (e.g., teacher, aid, researcher). Because the ages of compulsory schooling may vary, studies will be included where the participants were considered to be in their normal, day to day setting for primary/secondary education. Studies carried out in laboratories, nurseries, camps, after-school clubs or similar settings will be excluded. Studies carried out in tertiary institutions, regardless of the age or enrollment status of students, will be excluded. Studies with the majority of students being outside the ages of 3-18 will be excluded. Studies with fewer than 10 students in total will be excluded.

b. **Interventions:** This review is interested in interventions explicitly based on theories of metacognition and self-regulation because of their potential apparent in previous reviews. Therefore, studies without a stated theoretical basis in these areas will not be included, even if the interventions' "active ingredients," such as discussion, journaling, or drawing, may appear similar to those in metacognitive and self-regulatory interventions. Studies with interventions that simply train students in problem solving techniques without any reference to metacognition or self-regulation will not be included. Conversely, all studies encountered that explicitly link their interventions to metacognition or self-regulation in the title or the abstract will be included, with no prior judgements made about the potential mechanisms of effectiveness, or lack thereof. Interventions need to be in a form that a classroom mathematics teacher could implement during regular lessons and be intended to train or improve students' metacognition or self-regulated learning skills. Studies that refer to MC/SRL in passing, or that measure MC/SRL beliefs or skills, but do not actively train such skills, will not be included.

c. Comparators: Studies will be included that utilize a control or comparison group not receiving the metacognitive or self-regulatory intervention, but still receiving mathematics teaching appropriate to the grade or year level. Intervention groups should be compared against students studying similar mathematical content, so for example a study would be excluded if it included an "honors" class in the intervention group, but a standard-ability class in the comparison group. However, if the study included mixed-ability groups or classes in the intervention and the comparison groups, that would not preclude inclusion. Studies with no comparison group, such as within-group only designs, would be excluded. Studies with ITT (intention to treat), in which not all students received the condition to which they were allocated will still be included, unless an ITT analysis indicates substantial bias in the outcomes. Natural experiments will be excluded.

d. Outcomes: The main outcome of interest, effect on mathematics achievement, should be pre-specified as an outcome in the design of included studies, even where those studies may have additional outcomes, such as motivation or language proficiency. The measurement of mathematics achievement should be done in such a way as to enable the calculation of an effect size, if one is not given in the study report. Mathematics achievement should be understood as the skills normally assessed in standardised assessments of mathematics, such as problem-solving, description or manipulation of numbers or figures, recall or explanation of mathematical principles or techniques, or the application of mathematical techniques to address given scenarios. The review will exclude studies in which attitudes to mathematics are assessed but not the types of skills listed above. The assessments used to measure mathematical achievement should have been used previously outside of the study, and both intervention and control or comparison groups should take the same assessments. Studies in which a new assessment has been created specifically for the research, and no other assessment of mathematical achievement is used, will be excluded from the review.

Appendix 3. Full inclusion and exclusion criteria

Report Aspect	**Included**	**Excluded**	**Rationale and examples**
Dates	Published from January 2005 to December 2019.	The study was published before 2005 or after 2019.	The object is to update several reviews done a decade ago, with a special focus for the present review on metacognition in the mathematics domain.
Publication Language	English	The study is not available in English.	This review is aimed at readers of English, which is a global academic language.
Publication type	Scholarly and trade journals and magazines subject to peer review, conference proceedings, and theses and dissertations. All publications should be listed in common English-language databases of scholarly work.	Exclude: monographs, books, and book chapters; journals, magazines, or websites not subject to peer review; works not listed in common academic databases; unpublished studies or data sets.	The peer review process is considered essential as a quality-control feature. Books and chapters are likely to not be the primary form in which results of relevant studies are disseminated. A listing in a common scholarly database is considered necessary to ensure consistency in accessibility.
Study Aspect	**Included**	**Excluded**	**Rationale and examples**
Topic	Studies in education.	Any study not regarding teaching and learning, even if metacognition (or similar) and/or mathematics are mentioned. Also exclude studies about research design per se.	Studies in the disciplines of medicine, psychology, neuroscience, business, law etc. will be excluded.

Population/ Sample	Students aged 3 to 18, enrolled in primary/elementary and secondary schools. There should be at least 10 students total in the study.	The subjects of the intervention are NOT students aged 3-18 in primary or secondary settings, or there are fewer than 10 students total in the study.	Given the differences in assessments, accountability, and teaching context, research that focuses on post-secondary students is less likely to be useful for primary and secondary teachers. Students who are 17 but enrolled as normal university students would be excluded. Students who are 18 but still in secondary school would be included. Students in Honors/AP/IB courses, considered to be university level but not open to normal post-secondary students, would be included. With fewer than 10 students, the results of a study are not likely to be useful predictors for the effects of the intervention in future research or pedagogy.
Research Setting	Schools or other normal settings for teaching and assessment in mathematics.	The study was done in a laboratory or other setting in which teaching mathematics is not a normal activity. Also exclude settings in which mathematics teaching is done intensively as a primary or only activity, such as summer schools or tutoring clubs. Also exclude studies in which the intervention was delivered at home.	This criterion aims to improve the ecological validity of the research. Studies done in laboratories or outside of normal teaching settings would be less useful to regular classroom teachers. A “laboratory school,” or other normal educational setting that also conducts educational research would be included. Any setting where the intervention cannot be compared with “standard teaching” done in the same setting would be excluded.
Design	Experimental or quasi-experimental studies, in which at least one aspect of teaching is manipulated, and some type of comparison group is included.	The study is NOT an experiment or quasi-experiment, in which at least one aspect of teaching is manipulated, and some type of comparison group is included.	The target audience of teachers should see an estimate of the impact of the intervention beyond standard teaching. This excludes measurement, theoretical, validation, or purely observational studies. Natural experiments or regression-discontinuity designs will be excluded due to the difficulty in disentangling the effects of the intervention from maturation or other systemic changes.

Intervention / Treatment	Educational intervention designed to be different from "standard teaching," with a named focus on metacognition or self-regulation. Some detail and documentation on the intervention should exist, to ensure the possibility of replication by another teacher or researcher.	The intervention is not different from standard practice, does not have documentation/detail in existence, is not based on an explicit metacognitive or self-regulated learning approach. The intervention period is less than at least 2 hours or two sessions long. Also exclude if the manipulated element of the study is embedded into the outcome measure AND there is no further follow-up outcome measure.	Teachers interested in this review would want to know what concrete ways are to change standard teaching approaches to improve learning. Any study without some detail on specific teaching techniques would not be useful for the target audience. Some studies may not feature interventions with an explicit metacognitive (or similar) approach, yet relate metacognitive outcomes in a general way, but such studies would be excluded from this review.
Outcomes/ Evaluation	Assessments of mathematical achievement given in quantitative results using standardised tests, curriculum assessments, school examinations, or cognitive measures. Data sufficient to compute an effect size (standardised mean difference) should exist. Documentation and/or detail on the outcomes should exist sufficient for replication by teachers/researchers.	Mathematics learning or achievement was not a designed outcome. Outcomes were not measured in a structured and replicable way. Observational, qualitative, or non-numeric outcomes are used, or there is insufficient data to compute an effect size.	While there are multiple good purposes for teaching activities, the target audience of this review is teachers interested in raising mathematics learning and achievement as measured by common standards and assessment tools. Reporting an effect size for a given intervention allows teachers to estimate the level of additional benefit in using that intervention with their own students. If no detail or documentation exists on the assessments used, teachers may question the accuracy of such an estimate.

Appendix 4. Final searches executed for the systematic review.

Database searched and access point	Search String	Limits applied	Date	Returns	Number of Duplicates
British Education Index (EBSCOhost)	AB (meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND AB (math*) AND AB (treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)	Publication Date: 20050101-20191231	25 May 2020	26	0
Education Abstracts (EBSCOhost)	AB (meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND AB (math*) AND AB (treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)	Publication Date: 20050101-20191231	25 May 2020	169	25
ProQuest Dissertations and Theses Global	ab(meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND ab(math*) AND ab(treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)	Date: From January 01, 2005, to December 31 2019	28 May 2020	145	0
Educational Administration Abstracts (EBSCOhost)	AB (meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND AB (math*) AND AB (treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)	Publication Date: 20050101-20191231	1 June 2020	48	44

Database searched and access point	Search String	Limits applied	Date	Returns	Number of Duplicates
ERIC-US Department of Education (EBSCOhost)	AB (meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND AB (math*) AND AB (treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)	Date Published: 20050101-20191231	2 June 2020	395	176
Web of Science Core Collection	(TS=(meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND TS=(math*) AND TS=(treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)).	LANGUAGE: (English) Timespan: 2005-2019. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-SSH, ESCI.	3 June 2020	898	210
PsycINFO-American Psychological Association (EBSCOhost)	AB (meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND AB (math*) AND AB (treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)	Publication Date: 20050101-20191231	3 June 2020	535	385
ProQuest Social Sciences Premium Collection	ab(meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND ab(math*) AND ab(treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)	Date: From 2005 to 2019	3 June 2020	305	272
JSTOR	(ab:((meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND (math*) AND (treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)))	2005 to 2019	3 June 2020	11	11

Database searched and access point	Search String	Limits applied	Date	Returns	Number of Duplicates
Scopus (Science Direct)	ABS (meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND ABS (math*) AND ABS (treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)	PUBYEAR > 2004 A ND (EXCLUDE (PUBYEAR , 2020))	3 June 2020	762	525
Educational Research Abstracts Online (Taylor and Francis)	(meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND (math*) AND (treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)	There was no option to limit by date.	2 June 2020	216	91
ECO-Electronic Collections Online (OCLC FirstSearch)	(((kw: meta-cogn* OR kw: metacogn* OR kw: self-reflect* OR kw: self-regulat*)) and kw: math*) and (kw: treatment* OR kw: interven* OR kw: experiment* OR kw: control* OR kw: compar* OR kw: condition* OR kw: trial* OR kw: random* OR kw: allocate*)	yr: 2005-2019	19 June 2020	91	47
Applied Social Sciences Index & Abstracts (ProQuest)	ab(meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND ab(math*) AND ab(treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*)	Date: From 2005 to 2019	3 June 2020	37	37

Appendix 5. Sources considered for review and rejected, with reasons.

Source and Outlet	Practice Search Executed	Date	Items Returned	Reasons for Discarding
ArticleFirst (OCLC FirstSearch)	(((kw: meta-cogn* OR kw: metacogn* OR kw: self-reflect* OR kw: self-regulat*)) and kw: math*) and (kw: treatment* OR kw: interven* OR kw: experiment* OR kw: control* OR kw: compar* OR kw: condition* OR kw: trial* OR kw: random* OR kw: allocate*) and yr: 2005-2019	20 May 2020	8	Not able to download references in a suitable format for EPPI-Reviewer, only a few returns, and the returns may be included in the ECO (FirstSearch) search.
Electronic Theses Online Service	meta-cognition (any word) AND mathematics (any word); metacognition (any word) AND mathematics (any word); self-reflect (any word) AND mathematics (any word); self-regulate (any word) AND mathematics (any word).	10 March 2020	43	Not able to execute search as a single string with alternate terms or use wildcards. Not able to download citations as a batch.
Sage Journals	[[Abstract meta-cogn*] OR [Abstract metacogn*] OR [Abstract self-reflect*] OR [Abstract self-regulat*]] AND [Abstract math*] AND [[Abstract treatment*] OR [Abstract interven*] OR [Abstract experiment*] OR [Abstract control*] OR [Abstract compar*] OR [Abstract condition*] OR [Abstract trial*] OR [Abstract random*] OR [Abstract allocate*]] 2005 to 2019	21 May 2020	44	The majority of these returns would have been picked up by other searches, leading to excess duplicates. Based on re-running the search on 3 Dec. 2021, I estimate 1-2 well-fitting items may have been lost by discarding this search.
Directory of Open Access Journals	(Article Abstract) metacogn* AND math*; (Article Abstract) meta-cogn* AND math*; (Article Abstract) metacognition AND math*; (Article Abstract) meta-cognition AND math*; (Article Abstract) self-reg* AND math*; (Article Abstract) self-regulation AND math*; (Article Abstract) self-reflect* AND math*; (Article Abstract) self-reflection AND math*	25 May 2020	133 (maximum from one search with two search terms)	Not able to combine more than two terms in a single search. Using wildcards produces inconsistent results.

Source and Outlet	Practice Search Executed	Date	Items Returned	Reasons for Discarding
WorldCat Dissertations and Theses (FirstSearch)	(((kw: meta-cogn* OR kw: metacogn* OR kw: self-reflect* OR kw: self-regulat*)) and kw: math*) and (kw: treatment* OR kw: interven* OR kw: experiment* OR kw: control* OR kw: compar* OR kw: condition* OR kw: trial* OR kw: random* OR kw: allocate*) and yr: 2005-2019 and la= "eng"	12 June 2020	190	Not able to download batch citations in the correct file format for uploading to EPPI-Reviewer.
Australian Education Research Theses	NA	NA	NA	Not able to access.
International Bibliography of the Social Sciences (ProQuest)	ab(meta-cogn* OR metacogn* OR self-reflect* OR self-regulat*) AND ab(math*) AND ab(treatment* OR interven* OR experiment* OR control* OR compar* OR condition* OR trial* OR random* OR allocate*) Additional limits - Date: From January 01 2005 to December 31 2019	20 May 2020	23	Part of the Social Sciences Premium Collection. Not necessary to do a separate search.

Appendix 6. Codes used for data extraction for the narrative synthesis.

- Report Type
    - Journal Article
    - Dissertation/Thesis
    - Conference Paper
    - Technical Report/White Paper
    - Other (Describe)
    - Possible overlapping reports?
- Date of Publication
    - 2005
    - 2006
    - 2007
    - 2008
    - 2009
    - 2010
    - 2011
    - 2012
    - 2013
    - 2014
    - 2015
    - 2016
    - 2017
    - 2018
    - 2019
- Study Context
  This section is used to extract basic information about the study context.
    - Continent/Region
        - USA
        - Central America and Caribbean
        - South America
        - Africa
        - Western Europe
        - Eastern Europe/ Former USSR
        - Middle East
        - Subcontinent
        - East Asia
        - Oceania/Pacific Islands
        - Canada
    - Country
      The country/nation in which the research was performed.
        - Given explicitly
          Enter the country in info and select text in the pdf.
        - Inferred
          Enter the country in info and select text in PDF that implies the country.
    - Region, State, or City
      Enter the local region, state, or city for the research.
        - Given explicitly or inferred
          Enter the location in info and select text in the pdf.

    - Unclear
      The more specific location is not clear from the text.
  - Type of Community
    Urban, suburban, rural.
    - Urban
    - Metropolitian
      City centre and suburbs
    - Suburban
    - Rural
    - Unclear
  - School Level
    Select the appropriate level(s) of the school(s) where the research was completed.
    - Preschool/Nursery
      Serving children aged 3 to 5 or 6, but excluding levels considered "compulsory."
    - Kindergarten
    - Primary/Elementary
      Including children aged 6 or 7 to 10 or 11, or from the start of compulsory education to early adolescence.
    - Middle School/Jr. High/Lower Secondary
      Including children aged 10 or 11 to 13 or 14, or from early adolescence to the final stage of compulsory education.
    - High School/Upper Secondary
      Including students from ages 13 or 14 to 17 or 18, or the final stage of compulsory education.
    - Secondary School
  - School Type
    - Publicly funded and administered
      State schools, run with public money and managed by a governmental entity. Any time a "district" is mentioned, use this code.
    - Privately funded and administered
      Include traditional private schools.
    - Publicly funded, privately administered
      Include charter schools in the US or similar.
    - Head Start, Abbot, or other Preschool
    - Unclear/Multiple School Types
      How the school is funded and administered are unclear, or there are several types of schools.
  - Questions and Notes
- Student Sample
  Use this section to extract information about the student sample under study.
  - Selection of students
    - Random or partly/pseudo-random
    - Purposeful/matching/stratified
    - Convenience or chosen by someone other than researcher
    - Unclear selection
  - Sample numbers
    Enter the total numbers of students and teachers who participated in the study.
    - Total number of students sampled/allocated to groups.
      Enter the total number of students and highlight in the text.

  - Unclear number of students sampled/allocated to groups.
    The number of students sampled/allocated to groups is unclear, or there is a difference between the two. Explain in info and highlight in the text.
  - Total number of students assessed for outcomes in all groups.
    Enter the total number of students assessed for outcomes and highlight in the text.
  - Unclear number of students assessed for outcomes in all groups.
    The number of students assessed for outcomes in all groups is unclear. Explain in info and highlight in the text.
  - Total number of "teacher" participants.
    Enter the total number of teachers or other school employees (e.g. classroom assistants, interventionists, counselors) who participated in the intervention, including those who team taught or assisted external researcher(s). If the intervention was delivered by researcher(s) through pull-out sessions, enter "0". Highlight information in the text.
  - Unclear number of "teacher" participants.
    The text does not give a precise number of teachers or other school employees who participated in the intervention.
  - Total number of classes
    The total number of classes allocated to groups, offered the intervention, or assessed for outcomes.
  - Unclear number of classes
    The number of classes allocated to groups, offered the intervention(s), or assessed for outcomes is unclear, or there is a difference between these. Explain in info and highlight in the text.
  - Total number of schools
    Enter the total number of schools allocated to groups, offered the intervention, or assessed for outcomes.
  - Unclear number of schools
    The number of schools allocated to groups, offered the intervention, or assessed for outcomes is unclear, or there is a difference between these. Explain in info and highlight in the text.

- Ability level(s) or functional designation(s) of student participants
  Enter information about the pre-study ability designation or grouping of student participants. Check all that apply, enter info and/or highlight text from the pdf.
  - Low ability/disability/SEN/IEP/"at risk"
    While many researchers and theories distinguish between these, such distinctions may not be meaningful for the review outcomes. So, they are lumped together here.
  - Average or mixed ability
    Students are not designated as having high or low ability or specific learning-related or general disabilities.
  - High ability or gifted
  - Home language different from school language
  - Unclear ability level or functional designation
    There is not enough information in the text to clarify student participants' ability levels or designations prior to the study.
- Age(s) of students
  - 3
  - 4
  - 5

        - 6
        - 7
        - 8
        - 9
        - 10
        - 11
        - 12
        - 13
        - 14
        - 15
        - 16
        - 17
        - 18
        - 19
        - Unclear age
    - Grade level of students
      Enter the grade designation for students.
        - Grades(s) explicitly given
          Enter the grade(s) as given in the text in info, and highlight the text if appropriate.
        - Unclear grade level
    - Socio-economic status
      Enter information about parents' or families' income or education level. Select all designations that apply to some or all student participants in the study. Enter info and highlight text to illustrate coding.
        - Low income or FSM, or mother's education
        - Middle income
        - High income or mother's education
        - Mixed SES
        - Unclear Income level, FSM, or other social support
          The income level or FSM eligibility of all students in the study is unclear.
    - Sex, gender, and orientation
      Enter any information given about the sex, gender, and/or orientation of student participants. Select all that apply. Explain in info and highlight text in the PDF.
        - Co-ed or mixed sex/gender
          The group of all participants includes students of every sex and/or gender.
        - Girls only
        - Boys only
        - LGTBQI
        - Unclear sex, gender, or orientation
          There is not enough information in the text to determine the student participants' sex, gender, or orientation.
        - Discusses gender as a relevant factor
    - Ethnicity/Nationality
      Add these codes based on information from the text.
        - Ethnicity(ies)/nationality(ies) given
        - Mixed or multiple ethnicities or nationalities
        - Unclear ethnicity/nationality
    - Questions and Notes
- District, school, class, and teacher samples

- District or local area
    - Selection
        - Random or intended random
        - Purposeful
        - Convenience or chosen by someone other than researcher
        - Unclear
    - Numbers sampled
        - Mentioned explicitly
        - Inferred
        - Unclear
    - Assignment to groups
        - Randomised
        - Psuedo- or intended randomised
        - Purposeful/Matching
        - Naturally occuring/Intact groups
        - Unclear
        - Not used as unit of assignment
- Schools
    - Selection
        - Random
        - Purposeful
        - Convenience or chosen by someone other than researcher
        - Unclear
    - Numbers sampled
        - Mentioned explicitly
        - Inferred
        - Unclear
    - Assignment to groups
        - Randomised
        - Psuedo- or intended randomised
        - Purposeful/matching
        - Naturally occuring/Intact groups
        - Unclear
        - Not used as unit of assignment
- Classes
    - Selection
        - Random or intended random
        - Purposeful
        - Convenience or chosen by someone other than researcher
        - Unclear
    - Numbers sampled
        - Mentioned explicitly
        - Inferred
        - Unclear
    - Assignment to groups
        - Randomised
        - Psuedo- or intended randomised
        - Purposeful/matching
        - Naturally occuring/Intact groups
        - Unclear
        - Not used as unit of assignment

        - Teachers or other staff who deliver the intervention
            - Selection
                - Random or intended random
                - Purposeful
                - Convenience or chosen by someone other than researcher
                - Unclear
            - Numbers sampled
                - Mentioned explicitly
                - Inferred
                - Unclear
            - Assignment to groups
                - Randomised
                - Pseudo- or intended randomised
                - Purposeful/Matching
                - Naturally occuring/Intact groups
                - Unclear
                - Not used as unit of assignment
- Grouping and Allocation
    - Intervention group(s)
        - Number and description
            - 1
            - 2
            - 3
            - 4
            - 5
            - Unclear number of intervention groups
            - Temporary Holding Cell
    - Comparison group(s)
        - Number and description
            - 1
            - 2
            - 3
            - 4
            - 5
            - Unclear number of comparison groups
    - Method of allocation
        - Randomised
        - Random with Matching
        - Psuedo- or intended randomised
        - Naturally occuring
        - Matching
        - Choice of student or school staff
        - Purposeful
        - Unclear method of allocation
    - Unit of allocation
        - Individual student
        - Group of students
        - Individual Class
        - Group of classes
        - Individual school
        - Group of schools

        - Other
        - Unclear unit of allocation
    - Questions and Notes
- Reject at this stage?
    - Maybe (give reasons)
    - Yes (explain)
        - Homework is a major part of the SRL/MC training
        - Insufficient intervention details reported
        - Intervention focuses on teachers not students
          The intervention training is pre-planned and delivered for teachers, and there is no or little direct MC/SRL training for students, even if student outcomes are measured.
        - Outcome assessment embedded in intervention
        - Comparison/control also includes MC/SRL elements
        - Numeric outcomes not acceptable
          Either I cannot understand the outcomes (due to inaccurate, confusing, or insufficient reporting), or the outcomes that are reported cannot be used to generate ES.
        - No reliable outcome assessment
        - Intervention is not clearly based on MC/SRL training for students.
        - Intervention not at least 2 hrs (1 session) or at least 1 hr (2 or more sessions)
        - Outside date range
        - Report describes an entire curriculum, not an intervention
    - No (explain resolution, if any)
        - Need to Code
- Research Timeline
    - Length of intervention period
        - Given explicitly
        - Inferred
        - Unclear
        - 1 week or less
        - 2 weeks to 4 weeks
        - 5 to 12 weeks or 1 academic quarter
        - 13-18 weeks or 1 semester
        - More than 18 weeks or up to 1 academic year
        - More than 1 academic year
        - Unclear Units of Time
    - Given or Estimated Time in Weeks
        - 1 week
        - 2 weeks
        - 3 weeks
        - 4 weeks
        - 5 weeks
        - 6 weeks
        - 7 weeks
        - 8 weeks
        - 9 weeks
        - 10 weeks
        - 11 weeks
        - 12 weeks

    - 13 weeks
    - 14 weeks
    - 15 weeks
    - 16 weeks
    - 17 weeks
    - 18 weeks
    - 19 weeks
    - 20 weeks
    - 21 weeks
    - 22 weeks
    - 32 weeks
    - 36 weeks
    - 40 weeks
  - Time-slot for intervention sessions
    - During normal maths period
    - During a free/study period
    - During another subject period
    - Unclear
  - Frequency of intervention sessions
    - Given explicitly
    - Inferred
    - Unclear
    - Daily or 4-5x per week
    - 2-3x per week
    - Weekly
    - Bi-weekly
    - Monthly or 1x per 3 weeks
    - More than 1 month between sessions
  - Length of intervention sessions
    - Given explicitly
    - Inferred
    - Unclear
    - Up to 15 minutes
    - Up to 30 mins
    - Up to 1 hour
    - Up to 1.5 hours
    - Up to 2 hours
    - Up to 3 hours
    - Up to 4 hours
  - Number of intervention sessions
    - Given explicitly
    - Inferred
    - Unclear
    - 1-2 sessions
    - 3-5 sessions
    - 6-10 sessions
    - 11-15 sessions
    - 16-20 sessions
    - 21-25 sessions
    - 26-30 sessions
    - 31-35 sessions

        - 36-40 sessions
        - 41-45 sessions
        - 46-50 sessions
        - 51-55 sessions
        - 56-60 sessions
    - Time between pre-test (if any) and first post-test
      Only tests done after the end of the intervention will be considered as posttests for this review.
        - Given explicitly
        - Inferred
        - Unclear
        - No pre-test
    - Time between first post-test and delayed post-test
        - Given explicitly
        - Inferred
        - Unclear
        - No delayed post-test was done
    - Questions and Notes
- Intervention
    - Name and development of intervention
        - Unique or specific intervention name
        - Unclear or generic intervention name
        - "Branded" or adapted from previous interventions
        - Newly developed, not "branded" or directly researched elsewhere
        - Unclear intervention development or links to other research
    - Intervention Type
        - Structure problem-solving with MC prompts/questions (e.g. IMPROVE)
        - IMPROVE or similar plus other intervention components
        - IMPROVE, or similar, NOT a main component
    - Description of intervention
        - Highly detailed in report
          Multiple examples are given and there are sufficient details to replicate many or most parts of the intervention, or to adapt it to a new context.
        - Moderately detailed in report
          Some examples are given, and there is detail allowing a few parts of the intervention to be replicated or adapted, but some parts are not clear.
        - Low detail in report
          Much about the intervention is unclear, but the basic outline and materials are given.
        - Described elsewhere
          Use this code when other reports or supplementary materials provide further description of the intervention materials and procedures.
        - Includes Samples of Intervention Materials or Activities
          Such as lesson plans, blank worksheets, tasks, computer screens, graphs/images, etc.
        - Includes Samples of Student work or discussion
          e.g. written texts, transcripts of dialogue, drawings. These should show examples of how students interacted with the materials.
        - Includes examples of teachers' or intervention leaders' work or discussion
          Such as logs, diaries, feedback to students, transcripts of recorded lessons, etc.

        - Includes a schedule of intervention sessions/lessons
          Do not check this unless there is a description of different sessions or lessons, or if they only describe the sessions in general.
    - Samples of mathematical tasks included
        - From intervention group teaching only
        - From control group teaching only
        - From shared teaching for intervention and control.
        - From assessment tool(s)
        - No sample maths tasks included
    - Theoretical basis and rationale for intervention
        - Discussed in article/report
            - In high detail
            - In moderate detail
            - In minimal detail
        - Unclear
        - Discussed with students
            - Explicitly stated or inferred
            - Not discussed
            - Unclear
        - Discussed with teachers or staff who led the intervention
            - Explicitly stated or inferred
            - Not discussed
            - Unclear
            - Researcher was teacher/intervention leader
    - Tailoring to specific needs of the students sampled
      Based on a formal or informal assessment of the needs of the specific students at the study site.
        - Tailored to Mathematics needs
            - By researcher
            - By teacher or intervention leader
            - By software or intervention structure
            - By students' choices
            - Unclear
        - Tailored to SRL/MC needs
            - By researcher
            - By teacher
            - By software or intervention structure
            - By students' choices
            - Unclear
    - Mathematical Content Area
        - Basic numeracy (e.g. counting, cardinality, recognizing written and spoken numbers)
        - Place value, ordinality, rank, or decimals, number sense
        - Adding/Subtracting
        - Multiplying/dividing
        - Fractions, percentages, ratios, or proportions
        - Time
        - Money
        - Shapes (classifying and describing)
        - Units
        - Data collection, organization, and presentation; measurement

    - Probability and Statistics
    - Numberline and graphing
    - Algebra and/or functions, rate of change calculations
    - Geometry, Circles, or Trigonometry
    - "Word Problems"
    - Logic
    - Estimating
    - Calculus or pre-calc.
    - Equations
    - Exponents
    - Unclear maths content
    - No specific maths content
- Level of mathematics embeddedness
    - MC/SRL taught through mathematics content
    - MC/SRL taught alongside mathematics content
    - MC/SRL taught through other or no academic domain
- Leader(s) of intervention
    - Regular classroom teacher(s)
    - Other teacher(s) or school employee(s)
    - Researcher(s)
    - Student(s)
    - Technology-based
    - Unclear intervention leader
    - Researcher as teacher or school staff member
- Training for intervention leader(s)
    - Live sessions led by researchers or intervention developers.
    - Teacher led live sessions
    - Manual, lesson plans or materials, or activity guides
    - Videos
    - Other
    - Unclear training
- Fidelity checks
    - Carried out
    - Not carried out
    - Unclear fidelity checks
- Social situation/modality
    - Individual student
    - Individual student with teacher or other adult
    - Pair of students
    - Group of students
    - Whole class
    - Internet-based (without live video/audio)
    - Video/audio meeting
    - Unclear
- Stage(s) of SRL

  What task stage (if any) was the intervention intended to focus on?
    - Pre-task
    - During Task SRL/MC
    - Post-task SRL/MC
    - Multi-stage or general
- Activities and Strategies

        - Mathematics problem-solving
        - Learning/reviewing mathematics principles
        - Graphing, modelling with images, taking photographs, or colouring
        - Construction or manipulation of physical props
        - Unstructured mathematics exploration, task formulation or choice by students
        - Setting learning or achievement goals
        - Defining or planning for tasks or learning activities, strategy choice
        - Questioning, monitoring, or control of learning or task processes or performance
        - Providing or receiving/reviewing marks or feedback, error-correction
        - Self-evaluation or self-prediction (e.g. calibration) of knowledge/performance
        - Attributions for performance
        - Writing about thinking
        - Discussion (verbal or written)
        - Behavioral, attentional, time, and environment regulation
          Including help-seeking from others in the social environment.
        - Affective/motivational regulation
          Including self-efficacy and beliefs about ability.
        - Memory and "study" strategies, note-taking
        - Explaining task approaches/TA (thinking aloud)/ students teaching peers, recording audio reflections
        - Physical exercise, movement, or breathing
        - Use of play, games, music/sounds, drama, or humor
        - Mathematical language/terms and reading strategies
        - General Reading Strategies
        - Unclear Activities/Strategies
    - Materials
        - Books, papers, notecards, or screens with text
        - Pictures, Images, Diagrams, Posters
        - Pencils, pens, markers, crayons
        - Rulers, protractors/compasses, or other measurement/drawing tools
        - Paint, colored or craft paper, sitckers, glue, scissors
        - PCs, laptops, tablets, phones (personal devices)
        - Whiteboard, smartboard, or "multi-touch" device
        - Passive video or audio display, slide show
        - Sports equipment or balls
        - Board games, playing cards, dice, coins, or manipulables
        - Dolls or toy animals
        - Gum, candy, or other consumables/prizes
        - Other physical materials
        - Unclear materials
    - Cost
        - Per student cost (GBP)
        - Per class cost (GBP)
        - Other cost estimate
        - Unclear cost
    - Social Acceptability
        - Student responses
            - positive

      - negative
      - unclear/no mention
    - Teacher and other staff responses
      - positive
      - negative
      - unclear/no mention
    - Students Revised Coding for Social Acceptability
      - Student responses positive only
      - Student responses negative only
      - Student responses mixed
      - Student responses unclear
    - Teacher responses revised coding social acceptability
      - Teacher responses positive only
      - Teacher responses negative only
      - Teacher responses mixed
      - Teacher responses unclear
- Evidence Strength and Risk of Bias
  - Blinding/masking performed?
    - yes
    - no
    - Unclear
  - Low or no Selection Bias?
    - yes
    - no
    - Unclear
  - Is baseline equivalence demonstrated?
    - yes
    - no
    - unclear
  - Is there a BAU comparison?
    - yes
    - no
    - Unclear
  - Evaluated at level of allocation?
    - yes
    - no
    - Unclear
  - Only one mathematics skills assessment tool (or equivalent forms)?
    - yes
    - no
    - Unclear
  - Mathematics Skills Outcome Assessment Seperately Developed/Validated?
    - yes
    - no
    - Unclear
  - Low/balanced Attrition/Missing data?
    - yes
    - no
    - Unclear
  - Low Contamination or Cross-over?
    - yes

        - no
        - Unclear
    - No evidence of participants' beliefs affecting outcomes?
      This could be resentful demoralisation or Hawthorne effect.
        - yes
        - no
        - Unclear
    - High Fidelity of Treatment?
        - yes
        - no
        - Unclear
    - Avoids Other Confounds and/or Limitations?
        - yes
        - no
        - Unclear
    - Questions and Notes
- Ethics considerations
    - Was consent/assent for the study sought from participants or parents/guardians?
        - Students
            - Yes
            - No
            - Unclear
        - Teachers/Staff
            - Yes
            - No
            - Unclear
        - Parents/Guardians
            - Yes
            - No
            - Unclear
    - Was personal data protected?
        - Students
            - Yes
            - No
            - Unclear
        - Teachers/staff
            - Yes
            - No
            - Unclear
    - Were participants free from reprisals/negative consequences?
        - Students
            - Yes
            - No
            - Unclear
        - Teachers/Staff
            - Yes
            - No
            - Unclear

Appendix 7. Extracted outcomes and effect size calculations for meta-analysis.

Data extracted and calculations of primary-study effects for the posttest-only meta-analysis can be viewed at:

https://docs.google.com/spreadsheets/d/124LJzgffzejKdqAPydZaZMica-9BL2poH47bRZpZrHE/edit#gid=0

Appendix 8. Assuming group sizes when not reported, impact on effect sizes and their confidence intervals.

Example 1. In Kang (2010), the original student sample is 75, but only 62 were included in the analysis due to attrition, missing assessment data, and lack of sufficient attendance in the study sessions. The sizes of the intervention and control group are not given in the report. Table A below shows the impact on the effect size and confidence intervals of different group size assumptions. Although different group sizes have a negligible impact, the outcomes could still be problematic based on the author's description of key differences between those who left the study and those who remained, such as the former having better mathematics performance scores (p. 52).

Intervention group size	Comparison group size	Posttest Only Effect size	95% Confidence Interval Minimum	95% Confidence Interval Maximum
31	31	0.8671	0.3464	1.3878
24	38	0.8648	0.3316	1.398
38	24	0.8695	0.3361	1.403

Table A. Comparison of outcomes based on different assumptions about how the total analysed sample (N=62) from Kang (2010) was subdivided.

Example 2. In Kramarski, Weisse, and Kololshi-Minsker (2010), the student sample is 140 (72 males, 68 females) spread across four classes and two different schools. Outcome descriptives are reported only based on two subgroups of achievement (higher vs. lower), and there is no information about how these groups were determined. For the meta-analysis, I assume equal groups by condition and achievement level, but the impact of changing the group sizes is explored in Table B below. As shown, there is a substantial difference in the effect sizes and their confidence intervals with different assumed group sizes. Importantly, all estimated effects and confidence intervals are positive, with moderate to high effects. There is some indication, based on the study report, that the intervention and comparison groups differed at baseline, with the latter having a wider range of pre-test scores and a bigger difference between the lower and higher achievement groups. The overall effect size is mainly the result of large pretest to posttest gains (M = 55.81 to M = 76.79) in the lower achieving students of the intervention group. Gains in the higher achieving intervention students were smaller (M = 76.92 to M = 84.00), and they were negligible in both the lower (M = 47.50 to M = 48.21) and higher achieving (M = 80.71 to M = 79.56) comparison group students (SDs ranged from 14.22 to 27.16).

Intervention group sizes		Comparison group sizes		Posttest Only Effect size	95% Confidence Interval Minimum	95% Confidence Interval Maximum
Lower achievers	Higher achievers	Lower achievers	Higher achievers			
35	35	35	35	0.7242	0.3823	1.0662

40	40	30	30	0.7493	0.4032	1.0953
30	30	40	40	0.7015	0.3569	1.0462
25	45	45	25	0.9556	0.6059	1.3053
35	45	25	35	0.6683	0.3246	1.0121

Table B. Comparison of outcomes based on different assumptions about how the total analysed sample (N=140) from Kramarski, Weisse, and Kololshi-Minsker (2010) was subdivided.

Example 3. Schmitt (2013) reports an original sample size of 276 total, with 126 intervention and 150 control group students. Reported attrition is 5% for the intervention group and 19% for the control group (with 241 final students overall, p. 74), so for the meta-analysis I assume the two final groups to have 120 and 121 students respectively. However, there is also a reported missing data on the mathematics outcome of 12% at pretest and 17% at posttest (26% for change scores, p. 83), and this missing data includes attrition and other reasons students missed the tests. Missing data not covered in the attrition rate is not reported by group, and only original samples are given in the outcome tables. Table C. below shows the differences in posttest-only effect sizes based on post-attrition calculated samples, and samples with an additional 17% missingness, either equal or unequal by group. As shown, because the overall effect is very small, there is almost no impact of changing the overall sample or using equal or unequal group sizes.

Intervention group size	Comparison group size	Posttest Only Effect size	95% Confidence Interval Minimum	95% Confidence Interval Maximum
120	121	0.0764	-0.1762	0.329
100	100	0.0763	-0.2009	0.3536
120	80	0.0752	-0.2078	0.3582
80	120	0.0776	-0.2054	0.3606

Table C. Comparison of outcomes based on different assumptions about how the total analysed sample (N=241 or N=200) from Schmitt (2013) was subdivided.

Example 4. Shilo & Kramarski (2019) reported 824 total students within 32 schools, with each school having one teacher and one class involved in the study. The number of students for each group is not reported, and there is no discussion of attrition or missing data. Equal-sized groups are assumed for the meta-analysis, but the impact of unequal groups is explored below in Table D. With all assumed group sizes, the effects are positive for the intervention and moderate in size. Given that the comparison group used an alternate intervention, and there was only one teacher involved in the study per school, it is not likely there was demoralisation and attrition based on knowledge of group assignment. However, even if there had been unequal attrition within the two study groups, this would not have impacted the effect size estimate substantially, as seen in the last two rows.

Intervention group size	Comparison group size	Posttest Only Effect size	95% Confidence Interval Minimum	95% Confidence Interval Maximum
421	421	0.393	0.2566	0.5294
321	521	0.3773	0.2371	0.5175
521	321	0.4109	0.2705	0.5514
421	321	0.4029	0.2562	0.5496
321	421	0.3839	0.2373	0.5304

Table D. Comparison of outcomes based on different assumptions about how the total analysed sample (N=824) from Shilo & Kramarski (2019) was subdivided. As no attrition was reported, the final two rows assume an imaginary attrition of 100 students from the total sample, unequal by groups, which is not assumed for the meta-analysis.

Appendix 9. Interventions and designs of review studies

All included reports, alphabetised by author, with intervention timing, description, and sample, comparison groups description and sample, and allocation method. U = Unclear from report., Ref. = Reference, Per. = Intervention period, Freq. = Frequency, Sess. = Session length, Interv. = MC/SRL Intervention, Interv. N = Intervention sample, Comp. = Non-MC/SRL comparison, Comp. N = Comparison group sample. Alloc. = Allocation method and unit.

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Abdolhossini (2012)	12 sessions	U	U	“Experimental group”: Multi-stage MC/SRL training, focused on memory, cognitive strategies, self-consciousness, control, attitudes to mathematics, self-assessment and problem-solving	100	“Control group”: assumed to be BAU mathematics teaching, but no details reported.	100	Randomised by school, teacher, and class.
Abdullah, Halim, & Zakaria (2014)	10 weeks	U	U	“VStops”: Trained all MC/SRL stages, focusing on visual representation of word problems.	96	BAU mathematics teaching: “The conventional approach refers to a direct approach to teaching and giving explanations using the year five textbook” (pp. 167-8).	97	U

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Aminah, et al. (2018)	1 semester	U	U	“Metacognitive teaching-learning approach”: Trained all MC/SRL stages, focusing on discussion of task approaches, “think alouds,” and “metacognitive journal writing” (p. 50).	36	“Conventional teaching”: Assumed to be BAU mathematics teaching, but no details reported.	34	U
Arroyo et al. (2007)	3 or 4 days	daily	30 min	“Tutor Intervention”: Used the Wayang Outpost online mathematics learning “tutor.” Students worked through sets of problems and received feedback and multistage, MC/SRL prompts after every 6th problem.	36	“Tutor Control”: Students use the Wayang tutor without automatic MC/SRL prompts, though they could request “hints” about the solution procedure only.	40	“Pseudo-random” by student
						“No Tutor Control”: BAU mathematics teaching, without using the Wayang Tutor. No pretest and no posttest of the primary mathematics outcome. Not used for effect size calculation.	38	Matched by class

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Babakhani (2011)	16 weeks or 2 months	U	45 min	“Cognitive and meta-cognitive strategies training with self-instruction procedure” (p. 566): Multistage MC/SRL problem-solving approach with verbal “modelling” by teacher and peers.	28	BAU mathematics teaching: “control group that does not receive strategy instruction” (p. 566).	30	Block randomised by gender
Baliram & Ellis (2019)	4 weeks	4x per week	20 min	“Metacognitive practice and teacher feedback”: Post-stage MC/SRL training with written responses to MC prompts and individual- and group-directed feedback from teachers.	33	“Comparison group”: Began and/or ended class with a content-review. No reflective writing or teacher feedback based on reflections.	42	Randomised by class
Barrus (2013)	1 semester (half an academic year)	daily	20 min	“SRL E-Learning Modules”: MC/SRL training focused mainly on domain-level functioning, not task-level. Emphasised adaptive person-related beliefs, motivation, and goal-setting/ monitoring. Used prior to daily practice with ALEKS.	12	BAU mathematics teaching: “traditional classroom teaching mathematics with direct instruction” (p. i)	15	Students randomised to classes, then classes randomised to condition.
						ALEKS-based class without SRL e-learning modules: “self-paced, individualised Algebra instruction with a web-based, intelligent tutor” (p. i)	12	

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Bond & Ellis (2013)	4 weeks	4x per week	5 min	“Metacognitive Reflective Assessment”/ ”Experimental group I”: Post-task MC/SRL training. To close each lesson, students wrote “I learned” statements, discussed them with a peer (“Think alouds”), then revised their writing.	47	“No reflection group”/ ”Experimental group II”: Closed each lesson with a short, non-reflective review of the mathematics teaching.	48	Teachers randomised to groups, with two teachers per condition
						“Control”: Worked on a different curriculum unit.	46	
Bruce (2015)	8 months	1x per week	U	“Intervention”: Mainly pre-task MC/SRL training. Students conferred with teachers to set performance goals based on MAP test scores in reading and mathematics and engaged in personalised learning to address goals.	107	BAU mathematics teaching: Pupils participated in MAP tests without structured goal-setting.	129	Random, stratified by pupil demographics and ability levels

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Byrd (2019)	12 weeks	3x per week	10 min	“Immediate Elaborative Feedback Using Student Response Systems”: Intervention trained multistage MC/SRL skills like, goal-setting planning, monitoring, and reflection. Students also shared MC feedback on mathematics problems through the ClassFlow™ SRS.	12 or 13	BAU mathematics teaching: “The control group only received immediate corrective feedback with the SRS. The control group did not receive the self-regulation strategy instruction” (p. 46).	12 or 13	U
Chen & Chiu (2016)	6 weeks	1x per week	40 min	“Collaboration scripts”: Pupils worked in teams using a multi-touch digital device to create a geometric design based on a prompt. The treatment group received automatic MC prompts to regulate their collaboration within and between teams.	35	“Comparison group”: completed the design activity without collaboration scripts.	37	Matching by class

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Cleary, Velardi, & Schnaidman (2017)	U	1x-3x per week	25 min	"Self-Regulation Empowerment Program (SREP)": Coaches provided multistage MC/SRL training in small groups focusing on performance feedback, goal-setting and monitoring, and mathematical beliefs and motivation. Mathematics teaching was only 20% of SREP session time.	21	"What I Need (WIN)": The in-place, remedial mathematics sessions used direct instruction, problem-solving, and group and peer work, without any explicit MC/SRL focus.	16	Stratified randomisation by teacher
Collingwood & Dewey (2018)	4 weeks	3x per week	45 min	"Thinking Your Problems Away": Multistage MC/SRL training for regulating motivation and affect along with problem-solving strategies. The intervention used mindful breathing, humour, and the IMPROVE self-questions for problem-solving.	72	Wait-list control group with BAU mathematics instruction.	72	Student-level matching and randomisation

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Cornoldi et al. (2015)	2-3 months	1x per week	1 hr	“Training 1”: The multistage MC/SRL approach focused on working memory and executive function along with a structured approach to understanding, visually representing, and solving word-problems.	69	“Training 2”: Wait-list control.	64	Allocated by class, method unclear
Cross (2009)	10 weeks	U	U	“Argumentation Group”: Within small groups, students discussed and defended their problem-solving approaches, with teacher prompting, then reported to the whole class.	43	“Control group”: Teacher-centred, direct mathematics instruction with problem-solving and only incidental discussion. No argumentation or writing activities.	55	Block randomisation by teacher within the treatment groups, control group allocation unclear
				“Writing Group”: Students “produce[d] a written argument justifying their response” (p. 911) to a problem, with teachers providing informal feedback for revision.	51			

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
				"Argumentation and Writing Group": Students alternated between the two approaches.	62			
Desoete (2009)	2 weeks	2x-3x per week	50 min	Pupils were taught a structured approach to "metacognitive prediction" (i.e., calibration) and problem-solving. To create interest in MC/SRL skills, the intervention used stories about "Number Town" and the animals that live there (Appendix 1, p. 13).	33	"Control group": Active control, in which students completed the same tasks without MC/SRL training.	33	U
Dresel & Haugwitz (2008)	1 semester	1x-2x per month	45 min	"Attributional feedback condition (AC)": Students practised with *MatheWarp*, an online mathematics tutoring programme, and received performance feedback and attributions aimed at increasing effort and use of problem-solving strategies.	42	"in a *placebo condition* (PC), students worked with the learning software and received feedback about the correctness of their answers but no attributional feedback" (p. 7)	48	Randomised by class

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
				"Attributional feedback enriched with metacognitive control questions condition (AMC)": Students used *Mathewarp* and received feedback and attributions as above, along with completing worksheets aimed at multistage MC/SRL skills.	61			
Edwards (2008)	11 or 12 weeks	daily	15 min	"Reflective assessment"/ "Intact group 1": Students evaluated their own daily and weekly learning in short writing compositions based on prompts. Peer-teaching was used once a week. Teachers gave written feedback and adjusted instruction based on students' reflections. Both conditions used a newly adopted curriculum, *Ramp up to Algebra*	27	"Intact group 2": Same mathematics teaching with the new curriculum, without reflective assessment.	27	"arbitrarily assigned by the investigator to one of the intact groups" (p. 79).

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Falco (2008)	9 weeks	1x per week	30 min	"Skill-builders": Multistage MC/SRL training programme focused on self-efficacy, goal setting, help-seeking, time-management, affective regulation and several other general and task-specific learning strategies.	79	"Comparison classes"/ "wait-list control": BAU mathematics teaching and received intervention after the study period.	74	Block randomised by teacher and class
Finau et al. (2018)	8 months	U	U	"Cognitive Acceleration in Mathematics Education (CAME)": 16 lessons, adapted for the research site, targeted general MC/SRL skills and specific reasoning patterns through group work and discussion around context-rich and ill-posed tasks, while teaching to mathematics objectives.	219	BAU comparison using government curriculum and generally teacher-centred approach. They were required to cover the same topics as the intervention group.	119	U

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Ford (2018)	3 weeks	daily	30 min	"Metacognitive training (MT)": A structured approach to solving word problems, infused with multistage, MC/SRL questions, with modelling by the teacher.	18	"Problem-solving strategy (PSS)": Taught the same problem-solving approach without MC/SRL questions.	15	Randomised by class
Fößl et al. (2016)	9 days	daily	50 min	"Experimental group (E)"/ "video supported seamless learning": students worked in teams to complete game-boards of mathematics tasks to meet specific learning objectives. Students could watch learning videos, ask teachers questions, and get correctness feedback in order to earn the most points.	24	"Control group (C)": Taught by teacher A, the same as the experimental group teacher. "traditional face-to-face mathematics instruction" (p. 324).	23	Allocated by class, method unclear
						"Further control group 1 (FC1)": Taught by teacher B. "traditional face-to-face mathematics instruction" (p. 324).	25	
						"Further control group 2 (FC2)": Taught by teacher C. "traditional face-to-face mathematics instruction" (p. 324).	13	

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Heemsoth & Heinze (2016)	2.5 weeks	daily	45 min	"Error-centred condition": The intervention trained post-task MC/SRL by asking learners to reflect on why they made errors on assessment items. They were also instructed to correct the error and construct a new item on which a similar error could occur.	87	"Solution-centered condition": Students were exposed to solved items similar to ones on which they had made errors. They had to explain the solution and then revise their own answers.	87	Randomised by student within each class
Hughes et al. (2019)	U	U	40-50 min	"Self-Regulation Mathematical Writing Strategy": Multistage, MC/SRL training through structured approaches to writing word problem-solving explanations. The instructor modelled the strategy with "metacognitive talk".	18	BAU control: "the control group received business-as-usual instruction and practice on expository/ informational writing" (p. 192).	9	Randomised by class

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Jackson Jackson (2012)	2 weeks	3x per week	30 min	“High Self Regulated /High Communal Learning Context”: Multistage MC/SRL approach in which students worked together to solve problems based on a structured method. Other intervention elements included keeping a strategy record and using self-praise to improve mathematics-beliefs.	37	“Low Self Regulated /High Communal Learning Context”: Students were encouraged to work collaboratively but were not taught the MC/SRL problem-solving approach.	30	Randomised by class
				“High Self Regulated Learning /Low Communal Learning Context”: Students learned the MC/SRL approach to problem-solving but practised it individually.	27	“Low Self Regulated /Low Communal Learning Context”: Students were instructed to work individually, without the MC/SRL problem-solving approach.	36	

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Jacobse & Harskamp (2009)	2 weeks	2x per week	30 min	“Task stairs”/“Treatment group”: used a computer programme that includes problem-solving with metacognitive hints that students choose based on different strategy categories, feedback on correctness, and prompting to reflect on task approaches after solving a problem. Multistage MC/SRL training.	23	“Control group”: BAU mathematics teaching without using the computer programme.	24	Allocated by class, method unclear
Jitendra et al. (2015)	6 weeks	daily	50 min	“Schema-based instruction (SBI)”: A structured approach to problem-solving and metacognitive strategies that replaced two standard units on ratio/proportion and percent. Teachers explicitly modelled the approach for students with dialogue and discussion.	944	BAU mathematics teaching from textbooks that varied by study site. Structured problem-solving was sometimes used but explicit MC/SRL strategies were not.	942	Randomised by class, with only one class per participating teacher

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Kang (2010)	1.5 weeks	daily	15 min	“Treatment group”: Trained pre-task MC/SRL skills through goal-setting exercises. Students had to set a goal for the number of problems they would solve on a daily worksheet.	31	Waitlist control completing the same tasks without trained goal-setting.	31	Randomised by student within each class
Kramarski & Dudai (2009)	5 weeks	1x per week	45 min	“Group feedback guidance (GFG)”: Small groups of pupils uses an online forum to discuss mathematics tasks, while working with a structured problem-solving approach, guided by metacognitive questions (i.e., IMPROVE). For this condition the questions were tailored to encourage collaboration and peer feedback. Multi-stage MC/SRL.	32	“Control group (CONT)”: Worked on the same mathematical tasks using problem-solving strategies but without an explicit metacognitive focus.	36	Allocated by class, method unclear.

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
				“Self-explanation guidance (SEG)”: The same as group feedback guidance, but this time the focus was on pupils’ explanations for their own solutions, using the IMPROVE self-questions. Tasks were still completed in the online forum. Multi-stage MC/SRL.	32			
Kramarski & Friedman (2014)	U	U	4 hr	“Solicited Prompts”: Students worked collaboratively in pairs supervised by a research assistant to complete mathematics learning tasks in a multimedia programme. In this group, they were able to choose metacognitive hints based on the IMPROVE self-questions to assist with problem-solving. Multi-stage MC/SRL.	30	“Control Group”: Pairs worked with the multimedia programme and were encouraged, like the other groups, to use discussion and “thinking aloud” to solve problems. They did not have access to the MC prompts.	30	Individual students were randomly allocated to pairs

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
				"Unsolicited Prompts": Students worked in pairs on the same multimedia programme but were given IMPROVE-based prompts on an automatic schedule. Multi-stage MC/SRL.	30			
Kramarski & Gutman (2006)	5 weeks	4x per week	U	"E-learning supported with IMPROVE self-metacognitive questioning (EL+IMP)": Pairs of students completed online tasks embedded with IMPROVE metacognitive questions. Teachers modelled problem-explanations and students had to respond in writing to the MC questions. Multi-stage MC/SRL.	35	"E-learning without explicit support of self-regulation (EL)": Pairs worked with the online tasks with no explicit MC/SRL focus, but teachers did show how to explain problem solutions.	30	Randomised by class

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Kramarski, Weisse, & Kololshi-Minsker (2010)	4 weeks	4x per week	1 hr	"Metacognitive Support (MS) group": Pupils used IMPROVE metacognitive questions to structure their task approaches. Teachers modelled them, and pupils explained their task approach in writing. Multi-stage MC/SRL.	U	"Non Metacognitive Support (N_MS) group": Pupils did not use IMPROVE but did discuss their problem-solving strategies in class.	U	Randomised by class
Kramarski & Zoldan (2008)	3 months	3x per weeks	45 min	"Diagnostic errors approach (DIA)": Through group discussions and writing exercises, students diagnosed conceptual errors, and they explained the correct solutions to mathematics tasks. Mainly post-task MC/SRL.	32	"Control approach. The CONT students were not exposed explicitly to metacognitive instruction, but rather practised the learning material individually or in groups. A discussion with regard to the solution was held either in small groups or among the whole class" (p. 140).	27	Randomised by class
				"Improvement via self-questioning (IMP)": Students used IMPROVE self-questions to practise MC/SRL skills in mathematics tasks individually and in groups. Multi-stage MC/SRL.	26			

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
				"Diagnostic errors approach embedded within the IMPROVE approach (DIA+IMP)": A combination of the other two interventions. Multi-stage MC/SRL.	30			
Lee, Yeo, & Hong (2014)	6 weeks	1x per week	1 hr	"STARtUP (STARt Understand and Planning)": Multi-stage MC/SRL training was delivered using a visual and written guide to structured approach to solving "non-routine" word problems with self-questioning. There was also a focus on diagramming the problems.	31	"Comparison class": Normal teaching using "think-pair-share" without a focus on diagramming problems or intentional MC/SRL skills.	32	Allocated by class, method unclear

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Lestari & Jailani (2018)	U	U	80 min	“Collaborative learning embedded within metacognitive strategies (COLAB+META)”: Trained multistage MC/SRL skills through collaborative group work on “reasoning mathematics tasks” (p. 2), while using IMPROVE self-questioning to structure the problem-solving approach.	62	“Collaborative learning with no metacognitive strategies (COLAB)”: Learners completed the same tasks in groups with no explicit MC/SRL training.	60	U
Mandaci Şahin & Kendir (2013)	8 weeks	U	U	“Experimental group”: Multistage MC/SRL training with a detailed problem-solving approach based on self-questioning. Teachers modelled the approach and then students used it independently, while using reflective writing at each stage of the process. Students discussed different approaches to the same task.	39	“Control group”: Problem-solving was done in the “traditional” way, without MC self-questioning. The teacher demonstrated solving a problem, students practised on their own then shared their solutions and corrected errors.	36	U

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
McClelland et al. (2019)	8 weeks	2x per week	15-20 min	“Self-regulation-only (SR)”: Children played circle time games, such as “Red Light, Purple Light,” designed to foster during-task MC/SRL skills, such as behavioural regulation. Children had to “pay attention to, remember, and follow increasingly complex sets of rules through multiple exposure” (p. 4).	59	“Business-As-Usual (BAU) delayed intervention group”: Control classrooms generally did not include “self-regulation games” (p. 8). Focus on mathematics or literacy is unclear.	37	Block randomised by teacher
				“SR+”: Numbers and letters are used to cue specific actions in the games, to boost “print knowledge and phonological awareness” and “counting and cardinality” skills (p. 3).	61			

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Mevarech & Amrany (2008)	1 month	U	U	“IMPROVE”: A structured problem-solving approach with self-questioning trained multistage MC/SRL skills. The teacher explained the approach and its rationale, and pupils recorded their thoughts in writing while using the approach.	31	“Traditional” mathematics teaching: Pupils worked on the same mathematics learning tasks without an intentional focus on MC/SRL skills.	30	Randomised by class
Mevarech et al. (2010)	1 month	5x per week	U	“IMPROVE”: A structured approach with MC self-questioning was used to solve word problems with “consistent” and “inconsistent language” (p. 197), training multistage MC/SRL skills. Group discussion and pictorial representation of the tasks were also emphasised.	100	“Control group”: “Traditional” mathematics teaching, “with no explicit exposure to meta-cognitive instruction” (p. 198).	94	Randomised by class

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Morales (2016)	6 weeks	2x per week	30-45 min	“Treatment group”/”Writing in mathematics”: Students were taught a multistage, structured approach to solving word problems, which emphasised group discussion, writing to explain thinking, and explicit teaching of mathematical vocabulary and sentence structure. English learners were given special assistance to complete the writing activities.	35	“Control group”: BAU mathematics teaching, with problem-solving practice. No explicit MC/SRL strategies or mathematical writing activities.	32	Allocated by class, method unclear

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Motteram et al. (2016)	10 months	2x per week	30 min	“ReflectED”: Scripted weekly lessons trained general MC/SRL skills, but these lessons are not reported in detail. Classes were also expected to have at least one unscripted session per week where pupils reflected on learning in multiple disciplines and used EverNote to catalogue written, photographic, or audio records of their reflections, to which they and their teachers could refer back.	800	Control group with “usual teaching”: Not described in detail, but some less-systematic MC/SRL teaching was already used in some schools.	707	Randomised by class

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
O’Neal (2015)	6 weeks	daily	5 min	“Experimental group”/”Metacognitive writing”: Mainly post-task MC/SRL skills were trained. Students reflected on their learning and addressed a different prompt for reflective writing at the end of each class. Students were instructed to write full responses, but there was no indication that reflections were used to guide later teaching and learning.	21	“Control group”: At the end of each class, students completed a problem based on the day’s learning and wrote it in their composition books, with no explicit MC/SRL training.	18	Randomised by class
Ozsoy & Ataman (2009)	9 weeks	2x per week	40 min	“Metacognitive strategy instruction using problem solving activities”: Students were given multistage MC/SRL training focusing on structured problem-solving with self-questioning, discussion, and writing reflections on learning.	24	“Control condition”: BAU mathematics teaching with individual problem-solving, teacher explanations, and error correction of students’ own solutions.	23	Randomised by class

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Pappas Schattman (2005)	2 weeks	1x per 2 days	20 min	“Intervention group”: Multistage MC/SRL skills were trained using common objects like coins and playing cards. The small group leader and children took turns acting out an arithmetical operation and then explaining their approaches, with increasing complexity. The focus was on raising children’s awareness of their addition and subtraction strategies.	20	BAU Control Group: “Children in the control group did not receive any metacognitive training or participate in any mathematical activities outside of the classroom” (p. 43).	24	U

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Pennequin et al. (2010)	7 weeks	1x per week	1 hr	"Experimental group": A multistage, structured problem-solving approach focusing on metacognition, problem interpretation and graphical representation, goal-setting, planning, and monitoring/control strategies.	23	"Control group": "instead of focusing on metacognitive skills, children in this group were given the usual instructional and study-guidance support consisting of memory, reading, writing, and mathematical (arithmetic and geometry) activities. None of the children, in either the experimental or control group, was taught how to solve the actual problem" (p. 207).	23	Randomised by student
Perels, Dignath, & Schmitz (2009)	3 weeks	3x per week	U	"Experimental group"/ "intervention": Focused on multistage and general MC/SRL areas such as attitudes to mathematics, goal-setting, motivation, planning, maintaining concentration, and responding to mistakes. A structured approach for solving specific problems was a minor focus.	26	"Control group": Normal mathematics teaching without an explicit MC/SRL focus, as well as "mathematical problem-solving strategies (selection, segmentation and display formats) which were not connected to the mathematical contents of the learning unit."	27	Allocated by class, method unclear

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Riggs (2012)	3 months	2x per month	U	“Practice group”: Students practised calibrations by predicting their score on topics-based sections of mathematics assessments. Following the assessments, students reviewed their predictions and their actual scores on each topic. Mainly pre-task MC/SRL.	53	“No practice group”: Students did not do calibrations for the mid-way assessments but only the final assessment.	57	Randomised by teacher
Rizk, Attia, & Al-Jundi (2017)	3 weeks	3x per week	45 min	“Experimental group”: Multistage MC/SRL skills are covered, including planning, self-questioning, and monitoring by thinking aloud. Teacher modelling, small group work, and formative assessment were used to train the strategies.	20	“Control group”: BAU mathematics teaching with the “traditional method" (p. 110).	20	Randomised by school

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Sarette (2014)	1 school year	2x per week	55 min	“Experimental classroom”: A multistage MC/SRL curriculum was developed, aimed at regulation of attention and affect, setting and monitoring mathematics performance goals, improving working memory, and discussing and choosing individualised strategies. Structured problem-solving and multimedia games were also incorporated.	17	“Control classroom”: BAU mathematics teaching.	17	Allocated by class, based on teacher interest

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Schmitt (2013); Schmitt et al. (2015)	8 weeks	2x per week	30 min	“Intervention” / “Playgroups”: Within normal classrooms, children were led in large-group circle time games such as “Red Light, Purple Light,” which focused on behavioural regulation and progressed in complexity. Considered during-task MC/SRL training, the games required “working memory, attentional flexibility, inhibitory control” (p. 77). There was no explicit mathematics focus.	120	“Control group”: BAU “Head Start” preschool curriculum.	121	Randomised by class
Shamir & Lifshitz (2013)	U	U	U	“E-book containing metacognitive guidance (EBM)”: Children engaged individually with an ebook, *Grandfather’s Minibus*, designed to	26	“E-book but without metacognitive guidance (EB)”: Children worked with an ebook version without any MC/SRL prompts.	25	Randomised by student

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
				teach early numeracy and arithmetic concepts. In this condition, the e-book had embedded metacognitive prompts to encourage more strategic use of the ebook features. Considered during-task MC/SRL training.		Control group: “Regular kindergarten activities” without the ebook.	26	
Shilo & Kramarski (2019)	4 months	1x per week	U	“Experimental group”: Trained teachers in mathematical discourse and metacognitive theory, with self-questioning in a structured problem-solving approach. The self-questions are based on the IMPROVE model. Pre-structured lessons based on the approach were delivered to students.	412	“Control group”: Alternate intervention focusing on mathematical discourse and also using questions to structure problem-solving, but without an explicit metacognitive focus.	412	

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Sings Jenkins (2009)	10 weeks	1x per week	U	“Treatment group”: The researcher provided the classroom teacher with scripts for training 10 strategies for different “phases” of MC/SRL, such as goal-setting and planning, organising and remembering information, test-preparation, explaining problem approaches, and “self-consequating” based on learning behaviours and performance. Teachers were to use the strategies as they fit their normal teaching.	25	“Control group”: BAU mathematics teaching, but no details reported. Offered intervention materials after the study.	30	Allocated by class, based on teacher interest

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Tok (2013)	8 weeks	4x per week	40 min	"Treatment group"/ "Know Want Learn strategy (KWL)": A structured approach to understanding and solving word problems, using a graphical framework to organise students' prior knowledge, what is given and required in the problem, and what they learned by solving the problem. The strategy was intended to reduce mathematics anxiety and boost performance, and it is considered multistage MC/SRL training.	24	"In the control group, instruction involved students' reading the problem silently, solving the problem individually, the teacher's checking the answers, a volunteer student's solving the problem on the board, and the correction of missing points and mistakes" (p. 202).	31	Randomised by class

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Tominey & McClelland (2011)	8 weeks	2x per week	30 min	“Intervention”/ “Circle Time Games”: Children joined in pull-out sessions using self-regulation games with music and movement, such as “Red Light, Purple Light.” The rules of the games were changing and progressed in difficulty, training working memory, attention, and inhibitory control. Percussion instruments and other simple props were used. Children were also given a chance to lead the games. Considered during-task MC/SRL.	28	“Control”: Normal preschool/Head Start teaching, focused on building pre-academic skills and using free-play or outdoor play, without an explicit MC/SRL component.	37	Randomised by student within classes

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Tzohar-Rozen & Kramarski (2013)	5 weeks	2x per week	1 hr	“Affective self-regulation group”: Multistage training for dealing with negative affect during mathematics learning. Activities and strategies included classifying emotions, intentional relaxation, positive self-talk, and a structured approach to solving “authentic” problems. Pupils were given short questions or prompts to be used in different problem-solving stages.	54	“control group”: alternate intervention training a structured approach to “authentic” problems, without an affective regulation component.	53	Randomised by class or school
Tzohar-Rozen & Kramarski (2017)	5 weeks	2x per week	1 hr	“MA (meta-affective) group”: Multistage training for dealing with negative affect during mathematics learning. Activities and strategies included classifying emotions, intentional relaxation, positive self-talk, and a structured approach to solving “serial” problems. Pupils were given short questions or prompts to be used in different problem-solving stages.	54	“control group”: alternate intervention training a structured approach to “serial” problems, without an affective regulation or metacognitive component.	53	Randomised by school

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
				“MC (metacognitive) group”: Multistage training for problem-solving with questions and prompts to structure the approach. Very similar to IMPROVE.	63			
Ubuz & Erdoğan (2019)	5 weeks	4x per week	U	“MAN+META group”/ “problems supported with explicit metacognitive questions”: Students worked with a tangram (a set of seven geometric shapes) to learn a unit on polygons. They were able to manipulate the shapes freely at first and then completed worksheets, embedded with metacognitive questions, to solve more formal tasks. The metacognitive questions are similar to IMPROVE and constitute multistage MC/SRL training.	129	“MAN group”/ “problems not supported with explicit metacognitive questions”: Students completed the same tangram exercises and worksheets, without the embedded metacognitive questions.	91	Randomised by class

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
Vula et al. (2017)	1 month	U	U	“Metacognitive instruction for solving math word problems”: Pupils collaborated with peers to work through a multistage problem-solving approach with metacognitive questions based on IMPROVE and SOLVE IT!. The approach focused on mathematical language, making sense of the problem, and visual representation, in addition to other strategies.	126	“control classes in which they performed tasks without having been given any specific guidance, based exclusively on traditional methods and respective textbooks” (Abstract, p. 1).	137	U
Wang et al. (2019)	13 weeks	3x per week	35 min	“SR condition”: Tutors led pairs of students in Super Solvers, focusing on basic “facts,” word problem schema, and applying	23	“Control Condition”: BAU mathematics teaching, with in-place intervention sessions for some students.	23	Randomised by student

Ref.	Per.	Freq.	Sess.	Interv.	Interv. N	Comp.	Comp. N	Alloc.
				strategies. Games were used to teach fluency, and a “motivational system” was used to manage behaviour. The additional SR component included multi-stage MC/SRL training, such as goal-setting, planning, and progress monitoring, and promoting a “growth mindset” and perseverance.		“Base condition”: Super Solvers was used without the additional SR component.	23	
Wijaya et al. (2018)	U	U	20-30 min	“Intervention Program”: Based on the concept of opportunity-to-learn, pupils worked on ill-posed, “context-based tasks” using worksheets with embedded metacognitive questions and prompts that were faded out over the intervention period. Teachers also prompted reflective discussion. Considered multistage MC/SRL training.	144	“Control classes”/ “regular program”: “teachers used a teacher-centered approach in which they mainly explained and demonstrated how to solve tasks . . . tasks all had a camouflage context and explicitly mentioned the mathematical concepts related to the task” (p. 9).	155	Allocated by class, chosen by principal (i.e., head teacher)

Appendix 10. Activities and strategies in MC/SRL programmes.

Codes are given for MC/SRL -based “activities and strategies,” with illustrative quotes from included studies and explanations for how they support MC/SRL skills. Note that each strategy was used in other studies beyond the examples given here.

Code name (most to least common)	**Reference**	**Coded Text**	**Explanation and relationship to MC/SRL skills**
Questioning, monitoring, or control of learning or task processes or performance	Chen & Chiu (2016)	“Students in the collaboration script condition received these computerized prompts that sequenced their discussions and interactions in all DBL phases, whereas students without the collaboration scripts did not receive any prompts to structure their collaboration. It is supposed that scripting of the interactions during learning would function as a catalyst that prompts metacognitive processes, therefore ensuring the intended learning takes place” (p. 270).	Students in both study conditions worked with multi-touch devices to complete an activity relating to shapes and patterns. The intervention group received automated prompts throughout the activity for how to constructively share their work with others. In this case, the strategy is for monitoring and control of social discourse, which is expected to also improve individual MC/SRL skills.
	Lee, Yeo, & Hong (2014)	“In the introductory session, Pòlya’s problem-solving approach and the STARtUP scheme were taught explicitly to the experimental class. Students were shown a problem and were given four sets of question cards, each with a different color. These were related to the four stages in Pòlya’s approach. They were then asked to order the four groups of questions that they would ask themselves when solving the problem (e.g., questions such as ‘‘What are given?’’ and ‘‘What do I need to find?’’ fall under the first stage). This introductory activity served to activate students’ prior knowledge of a logical sequence amidst their idiosyncratic problem-solving approaches” (p. 471).	Inspired by Pòlya, the researchers designed a problem-solving approach with five steps, shown as points on a star (Fig. 2, p. 469). Pupils are guided to decipher the language of the problem, use a pictorial representation, choose an appropriate “heuristic” to solve it, start their plan, and implement repair strategies as needed. This scheme exemplifies an MC/SRL approach because it raises students’ awareness of their problem-solving moves, gives them multiple strategy options, and prompts them to be sensitive to how the chosen strategy is working, making adjustments as needed.

Mathematics Problem-solving	Byrd (2019)	“After accessing the account, the teacher shared the mathematical content to the students’ devices to allow students to answer a question. Students reviewed the question and submitted their answers, which appeared anonymously on the ActivPanel in the classroom. . . When the answers appeared on the screen, students sorted the answers into different sections to begin the mathematical discourse. The discourse included feedback regarding the answers that appeared on the screen. Hattie, Fisher, and Frey (2017) noted that the discourse provides opportunities for students to share the correct answer as well as voice agreement or disagreement with the answers” (p. 52).	In this study, the problems to be solved are likely simple ones that can be answered quickly using the Student Response System. Still, these responses are used as a basis for active discussion. Pupils can compare their answers with others’, and receive feedback on their approaches, while saving face in front of their peers. This could encourage a more objective perspective as students reach a “meta” level in their thinking. It could also clarify mathematical principles and present pupils with more problem-solving options to choose from in the future.
	Finau et al. (2018)	“Give out a copy of Worksheet 1 (the map of the island) to each pair/group. Explain there are three possible landing sites available (shown on the map as A, B and C), that the toxic waste sites are marked (as 1and 2) and that on the island there is a big lake that the robot cannot cross. You could suggest that the robot lands at A, or the class could agree on a landing point, or each group could choose their own landing point. Tell the class that you are not expecting them to be able to solve the problem yet. Give them two minutes to discuss in groups what problems there are and what information you need from me to overcome them” (Electronic Supplementary Materials, p. 2).	This quote, taken from the teacher’s guide for implementing the Cognitive Acceleration through Mathematics Education (CAME) lessons, shows a more complex approach to problem-solving. Students have to work in groups to address a task in a fictional situation by first identifying the necessary information and steps. While basic operations would be required, finding an optimal solution to the “realistic” task would likely involve explanations of thinking and conflict-resolution, opening up a space for metacognition. There would also be opportunities for peer- and self-regulation within the group exercise.

Defining or planning for tasks or learning activities; strategy choice	Dresel & Haugwitz (2008)	“As a first step in setting goals and planning, we evaluated prior knowledge with two open-ended questions (i.e., “What can I already do well?” and “What do I still have trouble with?”). We designed one open-ended question to stimulate actual planning and goal setting (“What will I do today? In what order?”) and two multiple-choice questions concerning the planning of an exercise strategy and a reference to the available MatheWarp clues that clarified the mathematical subject matter (e.g., “I start with the moderately hard problems,” “First, I read through the relevant clues”). Two further aspects actively supported the monitoring control process: (a) one instruction (i.e., “Ask yourself now and again whether your plan is still valid”) and (b) two multiple-choice questions, which are to be answered after the student has worked with MatheWarp for about 10 min and that we aimed to prompt procedure changes in response to unexpectedly good or poor exercise performance (e.g., “Starting now, I will only work on the hard problems,” “I will use the clues for help”)” (p. 8).	This example shows an approach to planning at both a domain and task level. Students worked with a computerised tutor to practise items related to the regular class curriculum, and the intervention prompts students to use the program in the most strategic ways. First, students have to identify their strengths and weaknesses and choose items of optimal difficulty for skill development. Further prompts relate to preparing for tasks within a specific curriculum area by reviewing basic principles and “clues.” This type of activity could spur students to be more independently metacognitive by synthesising knowledge of the task or domain with knowledge of themselves as thinkers to improve learning and performance.
	Jackson Jackson (2012)	“Organizing: Make a Strategy sheet and place it in your planner; keep it handy! Select the strategy that best organizes your thinking to solve the problem Make a chart or graphic organizer to display your thinking Draw a picture Use manipulatives (concrete methods of counting)” (pp. 79-80).	This example from the teaching materials for the MC/SRL training suggests ways learners can record and recall their own productive strategies for future use. These concrete approaches could help make explicit the development of one’s thinking, and they might especially appeal to teachers of young children. They also provide scaffolding for learners to be more independent in approaching a task, rather than relying on only taught strategies.
Self-evaluation or self-prediction (e.g., calibration) of knowledge/performance	Vula et al. (2017)	“After having solved "the equation" learners are instructed to have a look at the visual representation and to decide whether their solution makes sense. Learners are encouraged to reread the math world problem and to have a look at the diagram that represents the sentence (in the drawing) in order to make sure that the information is correctly presented. Noting down the relations between the numerical data (quantity) in the problem with symbols and drawings aids learners in all the steps described above” (p. 54).	Here, students are led to self-check their task solutions, rather than relying on only external feedback. To do so, students need to consider whether the solution fits their domain- and task-related schema, while also utilising the provided information, such as diagrams of the task. This implies a move from a mechanistic application of procedures to a more reflective approach that aligns well with MC/SRL theories.

	Desoete (2009)	“Each of the metacognitive sessions involved a metacognitive prediction strategy based on a combined approach with modelling by teachers and improving metacognitive knowledge and skills … The training was verbal in nature and focused on the prediction of task difficulty as well as on the tasks and problem-solving procedures themselves” (p. 442).	This study evaluated whether asking students to do pre- and post-task predictions of their performance (treatment) would lead to better achievement than simply practising on problems (control group). This activity, elsewhere referred to as “calibration,” shows students’ awareness of themselves and the task requirements and difficulty. This type of awareness is necessary but not sufficient for improved learning in MC/SRL models.
Discussion (verbal or written)	Aminah et al. (2018)	“The third component of MTLA is pair discussion, group discussion, class discussion. According to Vygotsky’s theory of social constructivism, learning meaningfully will occurs in a social context. When students interact with each other, they share information and suggestion to other members of the group. All members of the group believe that they need each other and receive feedback and they share their ways of thinking and their ways to solve problems to each other. By metacognitive scaffolding, then students construct their new insight, knowledge, and skills meaningfully. Like that, learning in small groups will motivate students to be able to overcome conflict and contradiction which arise while discussion happened, report and they construct a new and more appropriate knowledge” (p. 50).	While the description of how discussion is used in this study is limited, there is a clear connection to MC/SRL theory, since pupils are given the opportunity to distinguish between different ideas, make their own explicit, and to resolve cognitive dissonance. The social setting is considered crucial here since these processes might not be initiated in independent work.
	Falco (2008)	After students finish the worksheet, the counselor facilitates a discussion focusing on a few key questions, such as, “Do you feel comfortable asking questions during class?” and, “Can you think of a time when you have wanted to ask a question in class but didn’t?” and, “What prevented you from asking a question when you needed to?” The counselor can help students process their responses” (p. 74).	In this case, discussion is led by the intervention-facilitator to help students understand potential affective barriers to using SRL strategies. Implicit in this is a goal of showing that all learners experience challenges, and that students can choose to be self-regulated even when it makes them uncomfortable, or they can choose alternate help-seeking strategies to minimise their discomfort.

Providing or receiving/reviewing marks or feedback, error-correction	Baliram & Ellis (2019)	“The participating teacher de-identified the reflective cards and made them accessible to another geometry teacher (not affiliated with the participants) to provide the feedback. This procedure was done to avoid bias and to allow students to receive content-specific feedback. Brookhart (2008) proposed that immediate or slightly delayed feedback should be provided while students are still mindful of the learning goal, concept, or assignment. In this study, students received content-specific feedback within 3 to 5 days” (p. 99).	In this example, intervention group pupils solved sample problems on notecards and used metacognitive prompts to write short reflective statements about their current knowledge. Shortly after, they were provided with feedback, but the report lacks samples of students’ written reflections or the types of feedback given. Still, it is clear the feedback is intended to update students’ self-awareness and identify learning needs. MC/SRL models suggest students may need support to use feedback productively.
	Fößl et al. (2016)	“Continuous feedback from the teacher during the working phase – via incentives and comments – motivated students to correct their mistakes. An evaluation of the students’ control sheet of their individual working progress showed that overall 396 exercises needed to be corrected by the teacher during the working phase and that in fact 299 of them were corrected again by the students. . . While “only” 68.2 % of the exercises rewarded with 1 star were corrected, 85.4 % of the difficult “3-star-exercises” were corrected. Furthermore, many students mentioned in the interview that they liked “collecting stars” and that they felt more motivated to do their exercises” (p. 331-332).	The intervention in this study leveraged a game-based approach to learning a unit of curriculum. Students working in teams worked at their own pace to complete tasks, with different numbers of points awarded based on their difficulty. Students had the option of correcting their errors based on feedback from the teacher. While the competition and variable incentives encouraged students to focus on the “difficult” problems, these may or may not have aligned with individual students’ needs. Still, it would be hoped that the motivational scaffolds here would allow students to see the benefits of using feedback independently in the future, thus improving MC/SRL skills.
Explaining task approaches/TA (thinking aloud)/ students teaching peers, recording audio reflections	Sings Jenkins (2009)	“Strategy 10: Writing the Steps (Performance Phase) Introduce it: People who study how students learn math have found that when students can explain a process, it helps them to understand it better and be more successful using it. Sell it: Model for students how you can go through each step used solving a problem and explain what you did and why. Show them an annotated example in the textbook. Ask students to work with a partner to work two problems. Each student should work one of the problems independently first. After they have finished, they should take turns explaining what they did and why in each step to their partner. The partner should ask probing questions to help the student provide complete explanations” (p. 121).	This sample lesson plan guides teachers in modelling problem-solving explanations and showing how such explanations could be a useful check on knowledge. While the ability to explain an approach does not guarantee the approach will always work, the implication is that the act of explaining it will help learners clarify their thinking. Importantly, there is an appeal to research to demonstrate the rationale for the MC/SRL strategy, however not all studies in this review seem to have presented students with a research-based rationale.

	Bond & Ellis (2013)	“Two separate reflection strategies were combined to form the independent variable: a written “I Learned” statement and a verbal “Thinking Aloud” strategy. These reflective strategies are efficient ways for teachers to facilitate student reflection on what has been learned while finding out if their lesson objectives have been attained. During the last five minutes of the lesson, students in the Experimental Group I were asked to think about what they had learned during the class period and then to write a sentence that began with the phrase “I learned.” Students were then prompted by the teacher to talk about what they had written with another student, the “Think aloud” strategy, and finally to edit as appropriate their “I learned” statement” (pp. 229-30).	In this intervention, “thinking aloud” is combined with writing to help students produce a more accurate record of their state of knowledge. The activity is also expected to aid teachers’ understanding of students’ thinking and where they might need further instruction. While “think alouds” (TAs) have been used as assessment tools and are often conducted “online” as students work through tasks, here they are used “offline,” after completing the task, and function as a self-assessment. There have been some concerns that think alouds may not always accurately capture students’ strategies in use, and that the cognitive demands of explaining could interfere with task performance, but these concerns might not apply with offline use of TAs.
Graphing, modelling with images, taking photographs, or colouring	Pennequin et al. (2010)	“Each training session had a different goal: The first concerned mental and concrete representations of the problem: how to create mental images, diagrams, drawings, or graphs. The researcher encouraged the children to compare their representations of the problem” (p. 207).	Where drawing or graphing were used in the interventions, they were mainly either the focus of the mathematics curriculum itself or used to aid the interpretation of problems, as seen here. Although the description of how the images were used in this study is somewhat unclear, it is inferred that they represent students’ thinking, and, once made “concrete,” allow others to view and respond to this thinking.

	Motteram et al. (2016)	“Enabling children to reflect well takes time and is scaffolded using a number of tools. A colour-coded system is used for the children to quickly show how well they thought they had done in a particular task. Green, for example, is used by a child to show they have been successful at the task and blue if they believe they have mastered the topic. Yellow denotes they are struggling, and red is used to show that they are ‘stuck’. A set of pictures showing different feelings, like happy faces or rain clouds, is also provided along with a range of emotion words for the children to express their feelings about the learning process. The colour code, pictures and words are displayed in the classroom for children to quickly refer to. In the early stages of the process children fill in paper-based templates, which are pasted into their workbooks. When children start using the Evernote software on the iPads to record their reflections, they can either take pictures of the paper-based templates or take pictures of the images or words expressing emotion and add a typed comment. These comments can become very detailed” (p. 6).	Here, the images are used to represent the learning task itself, as well as the social and physical environment and pupils’ emotions relating to their learning. The symbolic system using different colours to indicate confidence levels might appeal especially to younger pupils or those with difficulties in verbal expression. It is not clear if the photographs include the children themselves or the classroom, but if so, this could offer a more objective representation of the learning situation to accompany pupils’ subjective responses. It is also not clear how the intervention prompted pupils to make use of their reflections in future learning.
Writing about thinking	Hughes et al. (2019)	“There were six major components of the [mathematical writing (MW)] strategy modeled during the intervention, (a) making sense of the word problem, (b) determining an appropriate plan to solve the problem, (c) drawing a representation of the problem, (d) explaining problem solving and reasoning, (e) concluding the paragraph by stating the answer, and (f) systematically checking all components of work” (p. 192).	This intervention took place in a mixed-ability, multi-grade writing class, not mathematics class. It was expected to be especially beneficial for students with special needs, who might struggle with the mathematical language to describe their reasoning (p. 188). As students record their thinking, they are guided through a structured approach to the task and helped to evaluate their solutions. Interestingly, the approach also includes drawing a picture of the problem, which could act as a scaffold to complete the written response.

	Morales (2016)	"During the first session the students worked together with the teacher to solve a word problem so that vocabulary was explicitly taught, students could receive feedback on strategies and skills learned, and ESOL strategies such as sentence frames, sentence structure, mathematical vocabulary, and new content vocabulary was explained to students. . . . The ELLs were able to write in their first language if they were not able to explain their answers in English. Additionally, they were allowed to verbally say their answers in Spanish and the teacher or another group member helped them translate those answers" (pp. 49-50).	In this example, writing in the intervention is used not only as a tool for reflection but to assist students in developing mathematical language. The teacher scaffolded this by covering vocabulary and sentence structure before asking students to synthesise these subskills into a larger text. Still, there appeared to be residual concerns that English learners would be unable to cope with the tasks demands. It is not clear whether allowing them to use their first language was planned or a last-minute concession.
Learning/ reviewing mathematics principles	Kramarski & Zoldan (2008)	"In all classes, the linear functions and graph representations unit was taught three times weekly over a period of 3 months . . . These topics were practiced by procedural skills such as computation (e.g., slope, intersection values) and by higher order skills that referred to problem solving and explanations. In particular, students were asked to draw conclusions and make algebraic generalizations on the basis of a given graph or algebraic expression (e.g., analyzing graphs, deciding whether certain mathematical expressions represent the given graphs)" (p. 139).	Although covering the basic mathematics principles of the target unit is also done in the control group, this knowledge is used in a unique way in the intervention group, as a foundation for "higher order" skills. One focus of the interventions in this study is careful consideration of task-related errors, which could reveal where the underlying concepts are weak. Previous research and theory have considered the intersection of domain-specific knowledge and MC/SRL skills.
	Cross (2009)	"John calculates the value of the slope of the line says m=1. Sue takes a look at the line and without any calculations says he is incorrect. a) Explain how Sue knows this without calculating the value of the slope herself" (pp. 927-8). "The general guidelines for the teachers included ensuring . . . ensuring that the students were making sense of the questions and developing better understandings of the mathematical content" (p. 911).	These two passages illustrate how the intervention activities in this study were meant to teach mathematics principles both implicitly, through the tasks, and explicitly if needed through teachers' commentary on them. Rather than requiring a solution, the tasks require students to explain their knowledge of basic concepts related to a mathematical topic, in this case graphing a line.
Behavioural, attentional, time, and environment regulation	Collingwood & Dewey (2018)	"Coping self-statements were based on cognitive behaviour modification (CBM) type responses identified by Kamann and Wong (1993), which are thought to encourage and help an individual cope during challenging situations" (p. 79).	In this example, the self-talk strategies are designed to enable learners to persist in the face of difficulties and negative affect. This type of regulation may enable the regulation of cognition per se.

	Cleary, Velardi, & Schnaidman (2017)	“The strategy learning and practice category, which represents approximately 60% - 70% of SREP sessions, provides explanation, modeling, and guided practice opportunities for students to use different strategies to directly enhance their learning (e.g., draw pictures when solving mathematics problems) or enable them to effectively manage their thoughts, behaviors, and learning contexts (e.g., self-quizzing, self-motivation, help-seeking, time management)” (p. 30).	Along with teaching other strategies, the intervention tutors in this study trained students in regulating their learning environment, such as by appropriately allocating time and other resources to the task. Getting help from others can be a productive MC/SRL tactic, but students may need guidance on when and how to do this to avoid becoming over-reliant.
Affective/ motivational regulation	Sings Jenkins (2009)	“Explain that sometimes rewards and consequences come from teachers, parents and other adults, but self-consequating is something each student does for himself or herself. Sometimes a reward is as simple as feeling good about ourselves when we know that we have done the right thing or being proud of an accomplishment. . . . Explain that there are also times when we are disappointed in ourselves such as when we meant to put in extra time studying for a test and didn’t. In this situation, the student may not earn the grade he/she would like, but that is a consequence that comes from the teacher. The student should also apply a consequence. Similarly to feeling proud when we do well, a consequence may be the bad feeling we have about doing less than our best. But students may apply a more concrete consequence by deciding to do some extra studying so they will be more prepared the next time instead of watching a show they had planned to” (p. 115).	In addition to regulating cognition and behaviour, included interventions also assisted pupils in regulating their emotions in productive ways. Here, the focus is not simply avoiding or relieving negative feelings related to performance, but rather focusing on the controllable inputs of performance and how emotions (positive and negative) can be leveraged to establish and sustain learning behaviours. It is notable that pupils are encouraged to view a reward or punishment system as something they can choose to enact for their own benefit, rather than passively accept. This is clearly a meta-level approach.
	Perels, Dignath, & Schmitz (2009)	“Then she presented possibilities for stopping bothersome thoughts . . . For example, instead of thinking “This task is too difficult for me”, they could say, “Maybe I can find a small part of the task which is interesting and easy to solve” (p. 23).	Here, students’ affect and motivation interact with their beliefs about themselves and the task, such as that some aspects of it may be more or less difficult or engaging. In order to use the suggested strategy, students need to recognize when they experience potential de-motivation and resolve to take action to improve their own affect, rather than seeing themselves as victims of it.

Setting learning or achievement goals	Bruce (2015)	“The independent variables for this study were the prescribed interventions of (a) explicitly teaching students individual goal-setting based on formative assessments; (b) having students critically assess areas for personal improvement; and (c) based on those goals and assessed areas, students setting and participating in activities to address those specific areas” (p. 31).	In this study, learners were assessed with the Measures of Academic Progress (MAP® tests) several times a year, and the goal-setting was aimed at improving these scores. In this case, MC/SRL processes are activated in considering what goals are appropriate, which sub-skills to focus on, and what learning activities will bring the most benefit.
	Kang (2010)	“Teachers kept giving students feedback on the goals they wrote on their worksheets. For example, there was a student who was very slow at answering questions (e.g., low processing speed or fine motor problems) and he typically was able to answer three to four problems. He often felt frustrated when asked to work on worksheets and therefore initially set a goal of zero. In this situation, the teacher stepped in and convinced the student that five or six might be a better goal for him” (p. 54).	Importantly, teachers not only give students performance feedback in this study. They also offer ongoing guidance to setting goals that will be motivating and achievable. With this student, the intervention leader encourages goals based on patterns of ability rather than the student’s current affect. Thus, the student’s MC/SRL skills were enlisted in the process of setting goals not only in implementing them.
Use of play, games, music/sounds, drama, or humor	McClelland et al. (2109)	“The games focus on the three aspects of EF (i.e., working memory, attentional or cognitive flexibility, and inhibitory control) and enable children to practice self-regulation in a classroom setting (i.e., children play the games in a large group, such as during circle time)” (p. 3) “In the SR+ version of the games, literacy (print knowledge and phonological awareness) and math (counting and cardinality and numerical knowledge) content is embedded into the cues children are asked to respond to. For example, when playing Red Light, Purple Light, instead of responding to colors, children are shown a circle with a number written on it. In addition to responding to the color (e.g., clapping when they see blue, stomping when they see orange), children are shown a number card and asked to perform the action as many times as represented on the card” (pp. 3-4).	Play was used in the included reports primarily with younger learners to build self-regulation skills, as here, or to add interest and capture children’s attention while addressing cognitive and metacognitive strategies. In this iteration of the “Red Light, Purple Light” intervention, two conditions are used: One with and one without an explicit focus on pre-academic skills. In both cases, the belief is that by building self-regulation skills, learners will be better prepared to benefit from later academic instruction.

	Cornoldi et al. (2015)	“Children listened to a short story describing an investigation conducted by a police inspector. While listening to the story, children looked at a picture showing the characters in the story with their physical features After hearing the story, the children were asked to remember the relevant information provided by one of the witnesses to the crime Then, they were guided to reflect on how working memory was involved in this activity and transfer this reflection to the context of math problem, where solvers should be able to identify and recall relevant information” (p. 430).	In this intervention, researchers used a story-based approach to teach skills important for mathematics learning, in this case, noting and remembering relevant information. While in other similar interventions the connections to academic learning are not always made explicit, here the belief is implied that children will learn more if they understand the rationale. The researchers also show a belief that memory skills can be conditioned, harking back to studies on metamemory a precursor to broader theories of metacognition.
Memory and "study" strategies, note-taking	Sings Jenkins (2009)	“Model how to develop a mnemonic by picking a set of steps or terms in the current lesson and creating a sentence that can be used to trigger the memory. Ask students to work with a partner to create their own. Point out to students that using a mnemonic is only useful if the mnemonic helps them connect to the information they are trying to remember and is easier to remember than original information” (pp. 119-120).	Rather than teaching children uncritical reliance on pre-made mnemonics, the intervention here encourages children to consider and use strategically what they know about their own memory. The researchers also demonstrate that not all MC/SRL strategies will be useful in every context. Learners should be adaptive and recognize when an approach is working and when to change it.
	Falco (2008)	“Then, explain that studying for math may take more time than other subjects because time must be spent solving problems. Remind students to use their time-management skills to give themselves enough time to do all their math homework – if they skip parts or rush through it, they won’t learn important concepts as well as they should. Explain that math builds on everything that is learned before, so it is important for students to learn each concept completely as they go” (p. 71).	While many of the interventions from the review focus at task-level MC/SRL skills, here the focus is at general learning in the mathematics domain. Students are taught to regulate their overall time and pacing in studying, and to evaluate and solidify what they already know before proceeding to a more challenging skill. Students will also need to decide what counts as “learn[ing] a concept completely,” and they may need to use external resources to check this. The instructions here also imply that learning goals are likely to trigger MC/SRL process more than performance goals.

Mathematical language/terms and reading strategies	Pennequin et al. (2010)	“The aim of the second session was to develop relevant strategies for solving mathematical problems; for example, ‘‘read the question several times’’, ‘‘cross out irrelevant information’’, ‘‘check the Calculations’’. The aim of the third session was to teach children how to identify the key words in order to interpret the problem correctly. Which are the most important words in the question? What are the interrogative pronouns (who, how much, etc.)? What is the unit of the response (Euros, number of years, number of marbles, etc.)? The aim of the fourth session was to teach children to identify the mathematical expressions: for example, ‘‘what is the remainder?’’ indicates subtraction; the word ‘‘to add’’ indicates addition” (p. 207).	In this activity, the intervention leader guides students in understanding the various components of a mathematical task statement. Students need to understand how morphemes and syntax represent different quantities, relationships, and operations. While reading a mathematical description or task statement might not always be considered an MC/SRL strategy, here it is because students need to reflect on what they do or do not understand from their reading of it. Considering the various components of the statement separately could be a useful repair strategy if understanding is lacking during the initial reading. On the other hand, students need to have a good grasp of the underlying mathematical concepts to make sense of the linguistic cues.
	Shamir & Lifshitz (2013)	“The number corresponding to each passenger is transmitted visually as an illustration demonstrating the order in which the passengers entered the minibus. Graphic symbols of those numbers and a finger pointing to each place appear simultaneously with the narrator’s vocalisation of the names of the ordinal positions” (pp. 39-40).	In this study, pupils interact with an e-book designed to teach basic numeracy and the concept of addition. This description shows how ordinal numbers are presented visually and linguistically within the same activity. Depending on the language, there may be different words, symbols, or pronunciations for ordinal numbers versus cardinal ones, although the report does not specify this for the language of the e-book. In the intervention group, pupils are given prompts to continue interacting with a page until they understand all the concepts, which requires metacognitive reflection on their current state of knowledge.
Unstructured mathematics exploration, task formulation or choice by students	Ubuz & Erdoğan (2019)	“The introduction was conducted either by teachers’ short presentation of the manipulative to the whole class as such “Today we will use 7-piece tangram consisting of seven different geometric shapes that you have learned before and let’s start to investigate them” and then students’ free play with manipulative” (p. 136). “students tried to construct different polygons by using two, three, or more pieces. Following that students discussed the definition of polygon within their groups. Thereafter, some groups shared their constructions and definitions with the whole class” (p. 137).	In this activity, students freely explore with the different shapes to uncover their properties and explore different combinations. Using manipulatives, the students can quickly form and test different hypotheses, and share their knowledge with others with concrete reference points. In doing so, students are likely to develop their metacognitive awareness. While the same manipulatives can also be used with predefined tasks, allowing them to be used in an unstructured way may support a shift from performance goals to learning goals.

	Heemsoth & Heinze (2016)	“Create a problem in which a similar error could have occurred. Solve this problem correctly. The problem construction in the last prompt required the students to understand the rationale behind their error. Thus, the second and the fourth prompt triggered the learners to reflect the rationale behind their errors” (p. 107).	In asking pupils to construct a problem similar to one in which they produced an error, the researchers are prompting them to better understand the source of their error. This is considered, at least partly, “unstructured exploration,” because there is no set formula for producing an acceptable response. They need to show a meta-level awareness of how problems require specific kinds of thinking and could expose misconceptions, so the focus is not on task performance per se. However, this awareness could also improve performance.
Construction or manipulation of physical props	Pappas Schattman, (2005)	“The trainer then put all the pennies back into the bucket and modeled a subtraction problem. She said, “Now let’s pretend that I have 3 candies (placed three pennies in one hand then onto the stage) and I eat 1 of the candies (pretended to eat the candy and put the penny back into the bucket). How many candies do I have left/now (counted the pennies on the stage and produced an answer)?” Next, say, “Tell me everything I did to get 2.” Again, each child was asked to describe what the trainer did to solve the problem and was prompted if the description was incomplete” (p. 36).	Here, the researcher establishes a physical and visual reference by using pennies to demonstrate basic mathematical operations like addition and subtraction. Children are prompted to take note of every action and to explain what they saw, indicating metacognitive awareness. Instead of using actual candies, which might arouse more natural interest, the intervention leader uses coins to represent candies, which could facilitate a shift to symbolic representation later on without a physical referent. Yet such a rationale does not seem to have been communicated to the children at this point.
	Sings Jenkins (2009)	“For this class, students are going to use an approach that doesn’t require any special materials other than notebook paper, pencils, and scissors. The process presented for organizing notes works to help study vocabulary terms, processes, and problem solving. Model the following process for students. First, fold a piece of notebook paper in half lengthwise (hotdog fold). Then cut strips into the top layer. . . . Lift the strip and on the paper underneath, write the definition of the vocabulary term or the part of the process required for the step number recorded on the top strip” (p. 117).	In this activity, students are taught a useful study strategy, relying on basic materials. They can use the constructed object to check their memory of important information. Rather than simply displaying all the notes together on one page, the format of this object encourages students to focus on one important idea at a time. There is also an aspect of gamification here, as students may wish to improve their score while quizzing themselves. Brief guidance is given on when to use the approach most effectively so that students can choose when and how to be strategic.

Attributions for performance	Falco (2008)	“The objective of the lesson is to increase students’ self-reflection and self-regulation by helping them become more aware of their mistakes and by providing them with strategies for correcting them. The lesson should also increase the accuracy of their self-evaluations and causal attributions; mistakes can be identified and corrected and are not necessarily caused by personal deficiency or low ability” (p. 72).	In MC/SRL models, “attribution” refers to how people interpret the causes of their performance on an outcome. Although only used in a few studies here, elsewhere attributional retraining has been extensively investigated. The passage here demonstrates that maladaptive attributions can prevent learners from making productive changes, but raising metacognitive awareness can help.
	Dresel & Haugwitz (2008)	“MatheWarp contained 142 attributional feedback statements in three primary classes: success effort (e.g., “This good result can be traced back to the high level of effort you gave”), success ability (e.g., “You are well versed in this topic”), and failure effort (e.g., “Your work was too cursory on this problem”). . . . The algorithm was adaptive in that it featured an internally programmed adjustment of the boundaries among the four success categories to individual achievement development” (p. 8).	In this intervention, attributions are prompted externally through a tutoring programme. As elsewhere, these prompts encourage learners to assign responsibility for performance to internal and malleable causes, rather than external, uncontrollable ones. The concept of “ability” used here is somewhat uncertain, since, especially in mathematics, learners are wont to consider ability an in-born trait. Here it seems to refer to something developed through effort. The computerised system makes attributional prompts easier to deliver, but it is not clear whether learners place the same faith in them as in their own attributions or those from another person.
Physical exercise, movement, or breathing	Collingwood & Dewey (2018)	“Exercises from the Smiling Mind App http://smilingmind.com.au/) were used at the start of each session. Focused breathing exercises have been found to have positive impact on maths and anxiety” (p. 79).	This intervention made use of multiple strategies for relieving mathematics-related negative affect. While strategies for managing affect sometimes rely on self-talk or changing pupils’ perspectives on the discipline, here the approach is through direct physical movement (i.e., breathing exercises). Negative affect can be distracting and stifle engagement with learning activities, and it may be especially frequent in mathematics learning. It is useful to present a variety of stress-relieving strategies pupils can choose from.

	Schmitt (2013)	“Red Light, Green Light, a widely used childhood game, was introduced in the second week of the intervention. First, children were instructed to play in the traditional way (i.e., red light means stop, green light means go). Then, additional colors were added that also mean stop and go (e.g., blue means go, yellow means stop). In subsequent weeks, rules were changed to add increased difficulty and complexity (e.g., opposites were introduced so blue means stop and yellow means go)” (p. 77).	Movement is used here as part of a game (i.e., “Red Light, Purple Light”) designed to train several SRL skills that could pave the way for academic learning. In the games, children need to attend to their movements and adjust them quickly in light of the changing rules of the game. The external movements display visually the children’s current state of knowledge about the game and may help them and their teachers develop better metacognitive awareness. It is not clear, however, whether the children were told about the purpose for the games.

Appendix 11. Moderator and sub-group coding for all included studies

Study name	Effect size	Standard error	Participants	Weeks of instruction (Dose)	Similar to IMPROVE (I) or NOT	Report type: journal (j), conference paper (c), dissertation (d), technical report (r)	Average age of participants in years
Abdolhossini (2012)	0.6032	0.144591837	200	18	NOT	c	14
Abdullah, Halim, & Zakaria (2014)	1.1072	0.154617347	193	10	NOT	j	9
Aminah, et al. (2018)	0.2549	0.240127551	70	18	NOT	j	15
Arroyo et al. (2007)	0.0789	0.229821429	76	1	I	c	15
Babakhani (2011)	0.4885	0.266632653	58	8	I	c	10
Baliram & Ellis (2019)	0.5601	0.237066327	75	4	NOT	j	14.147
Barrus (2013)	0.5385	0.352244898	39	18	NOT	d	15.7
Bond & Ellis (2013)	0.9186	0.186836735	141	4	NOT	j	10.567
Bruce (2015)	0.4252	0.132193878	236	32	NOT	d	11
Byrd (2019)	-1.0579	0.435867347	24	12	NOT	d	10
Chen & Chiu (2016)	0.0732	0.235867347	72	6	NOT	j	10.5
Cleary, Velardi, & Schnaidman (2017)	-0.3518	0.334362245	37	14	NOT	j	12
Collingwood & Dewey (2018	0.3679	0.168061225	144	4	I	j	8.09
Cornoldi et al. (2015)	0.1039	0.173647959	133	8	NOT	j	9
Cross (2009)	0.3569	0.157780612	211	10	NOT	j	14.5
Desoete (2009)	1.1396	0.265408163	66	2	NOT	j	8.5

Study name	Effect size	Standard error	Participants	Weeks of instruction (Dose)	Similar to IMPROVE (I) or NOT	Report type: journal (j), conference paper (c), dissertation (d), technical report (r)	Average age of participants in years
Dresel & Haugwitz (2008)	0.7943	0.180637755	151	18	I	j	11.7
Edwards (2008)	-0.4045	0.274923469	54	11	NOT	d	14.5
Falco (2008)	0.00	0.161785714	153	9	NOT	d	11.204
Finau et al. (2018)	1.03	0.120561225	338	32	NOT	j	12.5
Ford (2018)	-1.6934	0.407015306	33	3	I	d	14.5
Fößl et al. (2016)	0.8353	0.249311225	85	1	NOT	j	10.6
Heemsoth & Heinze (2016)	0.2247	0.152091837	174	2	NOT	j	13.02
Hughes et al. (2019)	0.2306	0.409438776	27	2	I	j	10.5
Jackson Jackson (2012)	0.1796	0.175765306	130	2	I	d	9.015
Jacobse & Harskamp (2009)	0.5035	0.296377551	47	2	I	j	11
Jitendra et al (2015)	0.0464	0.046071429	1886	6	I	j	12.667
Kang (2010)	0.8671	0.265663265	62	1	NOT	d	17.2
Kramarski & Dudai (2009)	0.4389	0.210612245	100	5	I	j	14
Kramarski & Friedman (2014)	0.9134	0.23372449	90	3	I	j	14
Kramarski & Gutman (2006)	0.4357	0.251709184	65	5	I	j	14.5
Kramarski & Zoldan (2008)	1.218	0.234209184	115	12	I	j	8
Kramarski, Weisse, & Kololshi-Minsker (2010)	0.7242	0.174464286	140	4	I	j	14.5

Study name	Effect size	Standard error	Participants	Weeks of instruction (Dose)	Similar to IMPROVE (I) or NOT	Report type: journal (j), conference paper (c), dissertation (d), technical report (r)	Average age of participants in years
Lee, Yeo, & Hong (2014)	0.2469	0.252959184	63	6	I	j	9.5
Lestari & Jailani (2018)	0.3296	0.182321429	122	1	I	c	13.9
Mandaci Şahin & Kendir (2013)	1.6118	0.265969388	75	8	I	j	10
McClelland et al. (2019)	-0.0397	0.188061225	157	8	NOT	j	4.25
Mevarech & Amrany (2008)	0.3834	0.258443878	61	4	I	j	16.7
Mevarech et al. (2010)	0.56	0.15	194	4	I	j	8.4
Morales (2016)	0.9965	0.259285714	67	6	I	d	9
Motteram et al. (2016)	-0.011	0.051607143	1507	40	NOT	r	9
O’Neal (2015)	-1.2112	0.349234694	39	6	NOT	d	16.949
Ozsoy & Ataman (2009)	1.9972	0.357168367	47	9	I	j	11.2
Pappas Schattman (2005)	0.4937	0.307295918	44	2	NOT	d	5.36
Pennequin et al. (2010)	1.4955	0.333571429	46	7	NOT	j	10.833
Perels, Dignath, & Schmitz (2009)	0.443	0.278112245	53	3	NOT	j	11.033
Riggs (2012)	-0.0214	0.190816327	110	12	NOT	d	11
Rizk, Attia, & Al-Jundi (2017)	2.8425	0.448316327	40	3	NOT	j	10
Sarette (2014)	0.1306	0.343341837	34	36	NOT	d	7.5
Schmitt (2013)	0.0764	0.128877551	241	8	NOT	d	4.308

Study name	Effect size	Standard error	Participants	Weeks of instruction (Dose)	Similar to IMPROVE (I) or NOT	Report type: journal (j), conference paper (c), dissertation (d), technical report (r)	Average age of participants in years
Shamir & Lifshitz (2013)	0.7111	0.247678571	77	2	NOT	j	5.88
Shilo & Kramarski (2019)	0.393	0.069591837	824	8	I	j	10.5
Sings Jenkins (2009)	0.6233	0.277244898	55	10	NOT	d	13
Tok (2013)	1.8599	0.324591837	55	8	I	j	11.5
Tominey & McClelland (2011)	0.3163	0.251989796	65	8	NOT	j	4.55
Tzohar-Rozen & Kramarski (2017)	0.7973	0.171122449	170	5	I	j	10
Ubuz & Erdoğan (2019)	-0.0759	0.136938776	220	5	I	j	12.5
Vula et al. (2017)	0.1756	0.123647959	263	4	I	j	8.5
Wang et al. (2019)	0.9993	0.312755102	46	13	I	j	8
Wijaya et al. (2018)	0.2122	0.116071429	299	2	I	j	13.8

References

(* indicates inclusions in the systematic review and meta-analysis)

*Abdolhossini, A. (2012). The effects of cognitive and meta-cognitive methods of teaching in mathematics. *Procedia - Social and Behavioral Sciences, 46*, 5894–5899. https://doi.org/10.1016/j.sbspro.2012.06.535

*Abdullah, N., Halim, L., & Zakaria, E. (2014). VStops: A thinking strategy and visual representation approach in mathematical word problem solving toward enhancing STEM literacy. *EURASIA Journal of Mathematics, Science & Technology Education, 10*(3), 165–174. https://doi.org/10.12973/eurasia.2014.1073a

*Aminah, M., Kusumah, Y. S., Suryadi, D., & Sumarmo, U. (2018). The effect of metacognitive teaching and mathematical prior knowledge on mathematical logical thinking ability and self-regulated learning. *International Journal of Instruction, 11*(3), 45–62. https://www.e-iji.net/dosyalar/iji_2018_3_4.pdf

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *The American Psychologist*, *63*(9), 839–851. https://doi.org/10.1037/0003-066x.63.9.839

*Arroyo, I., Ferguson-Walter, K., Johns, J., Dragon, T., Mehranian, H., Fisher, D., Barto, A., Mahadevan, S., & Woolf, B. (2007). Repairing disengagement with non-invasive interventions. *Frontiers in Artificial Intelligence and Applications, Volume 158: Artificial Intelligence in Education.* https://ebooks.iospress.nl/volumearticle/3523

Atkinson, K. M., Koenka, A. C., Sanchez, C. E., Moshontz, H., & Cooper, H. (2015). Reporting standards for literature searches and report inclusion criteria: Making research syntheses more transparent and easy to replicate. *Research Synthesis Methods*, *6*(1), 87–95. https://doi.org/10.1002/jrsm.1127

*Babakhani, N. (2011). The effect of teaching the cognitive and meta-cognitive strategies (self-instruction procedure) on verbal math problem-solving performance of primary school students with verbal problem- solving difficulties. *Procedia - Social and Behavioral Sciences, 15,* 563–570. https://doi.org/10.1016/j.sbspro.2011.03.142

Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher, 48*(4), 217–228. https://doi.org/10.3102/0013189X19848729

*Baliram, N., & Ellis, A. K. (2019). The impact of metacognitive practice and teacher feedback on academic achievement in mathematics. *School Science and Mathematics, 119*(2), 94–104. https://doi.org/10.1111/ssm.12317

Baumfield, V. (2006). Tools for pedagogical inquiry: The impact of teaching thinking skills on teachers. *Oxford Review of Education, 32*(2), 185–196. https://www.jstor.org/stable/4618653

Baumfield, V. M., Hall, E., Higgins, S., & Wall, K. (2009). Catalytic tools: Understanding the interaction of enquiry and feedback in teachers' learning. *European Journal of Teacher Education*, *32*(4), 423–435. https://doi.org/10.1080/02619760903005815

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, *50*(2), 248–287. https://doi.org/10.1016/0749-5978(91)90022-L

Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology, 41*(3), 586–598. https://doi.org/10.1037/0022-3514.41.3.586

*Barrus, A. (2013). *Does self-regulated learning-skills training improve high-school students' self-regulation, math achievement, and motivation while using an intelligent tutor?* [Doctoral dissertation, Arizona State University]. ProQuest Dissertations & Theses Global. http://www.proquest.com/pqdtglobal/docview/1353661903/abstract/CF952F8887514C7EPQ/1

Bigelow, K. M., & Morris, E. K. (2001). John B. Watson's advice on child rearing: Some historical context. *Behavioral Development Bulletin, 10*(1), 26-30. https://doi.org/10.1037/h0100479

Bishop, J. (2021, September 20). *Pragmatist philosopher John Dewey's theory of religion.* Bishop's Encyclopedia of Religion, Society and Philosophy. https://jamesbishopblog.com/2021/09/20/pragmatist-philosopher-john-deweys-theory-of-religion/

Bisset, J. (2020, November 18). *Guides: Library research support: Research skills: Systematic review support.* https://durham-uk.libguides.com/research_support/research_skills/systematic_reviews

Boekaerts, M. (1999). Self-regulated learning: Where we are today. *International Journal of Educational Research, 31*(6), 445-457.

*Bond, J. B., & Ellis, A. K. (2013). The effects of metacognitive reflective assessment on fifth and sixth graders' mathematics achievement. *School Science and Mathematics, 113*(5), 227–234. https://doi.org/10.1111/ssm.12021

Borkowski, J. G. (1992). Metacognitive theory: A framework for teaching literacy, writing, and math skills. *Journal of Learning Disabilities*, *25*(4), 253–257. https://doi.org/10.1177/002221949202500406

Borkowski, J. G. (1996). Metacognition: Theory or chapter heading? *Learning and Individual Differences, 8*(4), 391-402. https://doi.org/10.1016/S1041-6080(96)90025-4

Borkowski, J. G., Estrada, M. T., Milstead, M., & Hale, C. A. (1989). General problem-solving skills: Relations between metacognition and strategic processing. *Learning Disability Quarterly, 12*(1), 57–70. https://doi.org/10.2307/1510252

Brown, A. L. (1977). *Knowing when, where, and how to remember: A problem of metacognition.* ERIC. https://eric.ed.gov/?q=brown%2c+metacognition&ff1=autBrown%2c+Ann+L.&id=ED146562

Brown, A. L. (1980). Metacognitive development and reading. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education* (pp. 453-482). Lawrence Erlbaum.

Brown, A. L. (1994). The advancement of learning. *Educational Researcher*, *23*(8), 4–12. https://doi.org/10.3102/0013189X023008004

Brown, A. L. (1997). Transforming schools into communities of thinking and learning about serious matters. *American Psychologist, 52*(4), 399–413. https://doi.org/10.1037/0003-066X.52.4.399

*Bruce, K. L. (2015). *A study of the impact of individual student goal-setting based on formative assessment on student achievement* [Doctoral dissertation, University of St. Francis]. ProQuest Dissertations & Theses Global. http://www.proquest.com/pqdtglobal/docview/1735491490/abstract/260B80D4D41E4E07PQ/1

Bryman, A. (2008). *Social Research Methods*. Oxford University Press.

Burke, J. F., Sussman, J. B., Kent, D. M., & Hayward, R. A. (2015). Three simple rules to ensure reasonably credible subgroup analyses. *BMJ*, *351*, h5651. https://doi.org/10.1136/bmj.h5651

*Byrd, L. A. (2019). *The effects of immediate elaborative feedback using student response systems on the mathematics achievement of fifth-grade students* [Doctoral dissertation, Mercer University]. ProQuest Dissertations & Theses Global. http://www.proquest.com/pqdtglobal/docview/2320997140/abstract/FEC5AD0AC8CF4B12PQ/1

Callahan, L. G., & Garofalo, J. (1987). Metacognition and school mathematics. *The Arithmetic Teacher, 34*(9), 22-23.

Campbell Collaboration. (2019, January). *Campbell systematic reviews: Policies and guidelines, version 1.4.* http://doi.org/10.4073/cpg.2016.1

Cardelle-Elawar, M. (1990). Effects of feedback tailored to bilingual students' mathematics needs on verbal problem solving. *The Elementary School Journal, 91*(2), pp. 165-175.

Cardelle-Elawar, M. (1995). Effects of metacognitive instruction on low achievers in mathematics problems. *Teaching and Teacher Education, 11*(1), 81–95. https://doi.org/10.1016/0742-051X(94)00019-3

*Chen, C.-H., & Chiu, C.-H. (2016). Collaboration scripts for enhancing metacognitive self-regulation and mathematics literacy. *International Journal of Science and Mathematics Education, 14*(2), 263–280. https://doi.org/10.1007/s10763-015-9681-y

Cherry, K. (2021, August 16). *What were functionalism and structuralism?* Very Well Mind. https://www.verywellmind.com/structuralism-and-functionalism-2795248

Chong, S. W., Collins, N. F., Wu, C. Y., Liskaser, G. M., & Peyton, P. J. (2016). The relationship between study findings and publication outcome in anesthesia research: A retrospective observational study examining publication bias. *Canadian Journal of Anesthesia/Journal Canadien d'anesthésie, 63*(6), 682–690. https://doi.org/10.1007/s12630-016-0631-0

*Cleary, T. J., Velardi, B., & Schnaidman, B. (2017). Effects of the Self-Regulation Empowerment Program (SREP) on middle school students' strategic skills, self-efficacy, and mathematics achievement. *Journal of School Psychology, 64*, 28–42. https://doi.org/10.1016/j.jsp.2017.04.004

Cobb, P., Gravemeijer, K., Yackel, E., McClain, K., & Whitenack, J. (1997). Mathematizing and symbolizing: The emergence of chains of significance in one first-grade classroom. In D. Kirshner & J. A. Whitson (Eds.), *Situated cognition: Social, semiotic, and psychological perspectives,* 151-233. Lawrence Erlbaum Associates, Publishers.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155. https://doi.org/10.1037/0033-2909.112.1.155

Cole, R., Haimson, J., Perez-Johnson, I., & May, H. (2011, September). *Variability in pretest-posttest correlation coefficients by student achievement level* [NCEE 2011-4033]. US Department of Education, Institute for Education Sciences. https://ies.ed.gov/ncee/pubs/20114033/pdf/20114033.pdf

*Collingwood, N., & Dewey, J. (2018, September). "Thinking your problems away": Can maths interventions be developed to address both the academic and affective aspects of learning in primary aged children? *Educational & Child Psychology*, 76–92. https://shop.bps.org.uk/educational-child-psychology-special-issue-september-2018-research-in-schools

*Cornoldi, C., Carretti, B., Drusi, S., & Tencati, C. (2015). Improving problem solving in primary school students: The effect of a training programme focusing on metacognition and working memory. *British Journal of Educational Psychology, 85*(3), 424–439. https://doi.org/10.1111/bjep.12083

*Cross, D. I. (2009). Creating optimal mathematics learning environments: Combining argumentation and writing to enhance achievement. *International Journal of Science and Mathematics Education, 7*(5), 905–930. https://doi.org/10.1007/s10763-008-9144-9

Davidson, J. E., Deuser, R., & Sternberg, R. J. (1994). The role of metacognition in problem solving. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (p. 207–226). The MIT Press.

de Boer, H., Donker, A. S., Kostons, D. D. N. M., & van der Werf, G. P. C. (2018). Long-term effects of metacognitive strategy instruction on student academic performance: A meta-analysis. *Educational Research Review, 24*, 98–115. https://doi.org/10.1016/j.edurev.2018.03.002

de Boer, H., Donker, A. S., & van der Werf, M. P. C. (2014). Effects of the attributes of educational interventions on students' academic performance: A meta-analysis. *Review of Educational Research, 84*(4), 509–545. https://doi.org/10.3102/0034654314540006

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review, 28*(3), 425–474. https://doi.org/10.1007/s10648-015-9320-8

*Desoete, A. (2009). Metacognitive prediction and evaluation skills and mathematical learning in third-grade students. *Educational Research and Evaluation, 15*(5), 435–446. https://doi.org/10.1080/13803610903444485

Desoete, A., Roeyers, H., & De Clercq, A. (2003). Can offline metacognition enhance mathematical problem-solving? *Journal of Educational Psychology, 95*(1), 188-200. https://doi.org/10.1037/0022-0663.95.1.188

Dewey, J. (1896). The reflex arc concept in psychology. *Psychological Review*, *3*, 357–370. https://brocku.ca/MeadProject/Dewey/Dewey_1896.html

Dewey, J. (2019). Moral principles in education. In *Moral Principles in Education and My Pedagogic Creed by John Dewey : With a Critical Introduction by Patricia H. Hinchey* (pp. 3-32). Myers Education Press.

Dewey, J. (2019). My pedagogic creed. In *Moral Principles in Education and My Pedagogic Creed by John Dewey : With a Critical Introduction by Patricia H. Hinchey* (pp. 33-50). Myers Education Press.

Dhuey, E., Figlio, D., Karbownik, K., & Roth, J. (2019). School starting age and cognitive development. *Journal of Policy Analysis and Management, 38*(3), 538–578. https://doi.org/10.1002/pam.22135

Digiacomo, G., & Chen, P. P. (2016). Enhancing self-Regulatory skills through an intervention embedded in a middle school mathematics curriculum. *Psychology in the Schools*, 53(6), 601–616. https://doi.org/10.1002/pits.21929

Dignath, C. & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school Level. *Metacognition and Learning, 3*(3), 231–264. eric. https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ817558&site=ehost-live

Dignath, C., Büttner, G., & Langfeldt, H.-P. (2008). How can primary school students learn self-regulated learning strategies most effectively?: A meta-analysis on self-regulation training programmes. *Educational Research Review, 3*(2), 101–129. https://doi.org/10.1016/j.edurev.2008.02.003

Donker, A. S., de Boer, H., Kostons, D., Dignath van Ewijk, C. C., & van der Werf, M. P. C. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review, 11*, 1–26. https://doi.org/10.1016/j.edurev.2013.11.002

Douglas, D., & Attewell, P. (2017). School Mathematics as Gatekeeper. The Sociological Quarterly, 58(4), 648–669. https://doi.org/10.1080/00380253.2017.1354733

*Dresel, M., & Haugwitz, M. (2008). A computer-based approach to fostering motivation and self-regulated learning. *Journal of Experimental Education, 77*(1), 3–18. https://doi.org/10.3200/JEXE.77.1.3-20

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*(10), 1040–1048. https://doi.org/10.1037/0003-066X.41.10.1040

Education Endowment Foundation. (2018, July). *Sutton Trust-EEF Teaching and Learning Toolkit & EEF Early Years Toolkit: Technical appendix and process manual (Working document v.01).* https://educationendowmentfoundation.org.uk/public/files/Toolkit/Toolkit_Manual_2018.pdf

Education Endowment Foundation (n.d.). *Teaching and Learning Toolkit: Metacognition and self-regulation.* https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit/metacognition-and-self-regulation

*Edwards, T. G. (2008). *Reflective assessment and mathematics achievement by secondary at-risk students in an alternative secondary school setting* [Doctoral dissertation, Seattle Pacific University]. ProQuest Dissertations & Theses Global. http://www.proquest.com/pqdtglobal/docview/304801373/abstract/AC942B603ED94AD0PQ/1

Efklides, A. (2009). The role of metacognitive experiences in the learning process. *Psicothema, 21*(1), 76-82. https://www.psicothema.com/pdf/3598.pdf

Efklides, A., Samara, A., & Petropoulou, M. (1999). Feeling of difficulty: An aspect of monitoring that influences control. European Journal of Psychology of Education, 14(4), 461–476. https://doi.org/10.1007/BF03172973

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (1994). The gender gap in math: Its possible origins in neighborhood effects. *American Sociological Review, 59*(6), 822–838. https://doi.org/10.2307/2096370

Ergen, B., & Kanadli, S. (2017). The effect of self-regulated learning strategies on academic achievement: A meta-analysis study. *Eurasian Journal of Educational Research, 17*(69), 55–74. https://doi.org/10.14689/ejer.2017.69.4

*Falco, L. D. (2008). *"Skill-Builders": Enhancing middle school students' self-efficacy and adaptive learning strategies in mathematics* [Doctoral dissertation, The University of

Arizona]. ProQuest Dissertations & Theses Global. http://www.proquest.com/pqdtglobal/docview/304683933/abstract/3A16A51039FF4583PQ/1

*Finau, T., Treagust, D. F., Won, M., & Chandrasegaran, A. L. (2018). Effects of a mathematics cognitive acceleration program on student achievement and motivation. *International Journal of Science and Mathematics Education, 16*(1), 183–202. https://doi.org/10.1007/s10763-016-9763-5

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

Flavell, J. H. (2000). Development of children's knowledge about the mental world. *International Journal of Behavioral Development*, *24*(1), 15–23. https://doi.org/10.1080/016502500383421

Flavell, J. H., Friedrichs, A. G., & Hoyt, J. D. (1970). Developmental changes in memorization processes. *Cognitive Psychology, 1*(4), 324–340. https://doi.org/10.1016/0010-0285(70)90019-8

*Ford, D. J. (2018). *The effects of metacognitive training on algebra students' calibration accuracy, achievement, and mathematical literacy* [Doctoral dissertation, Old Dominion University]. ProQuest Dissertations & Theses Global.

*Fößl, T., Ebner, M., Schön, S., & Holzinger, A. (2016). A field study of a video supported seamless-learning-setting with elementary learners. *Journal of Educational Technology & Society, 19*(1), 321–336. https://www.jstor.org/stable/jeductechsoci.19.1.321

Gafoor, K. A., & Sarabi, M. K. (2015, December 21-22). *Relating difficulty in school mathematics to nature of mathematics: Perception of high school students from Kerala* [paper presentation]. National Conference on Mathematics Teaching-Approaches and Challenges, Regional Institute of Education (NCERT). https://eric.ed.gov/?id=ED566898

Garofalo, J. (1986) Metacognitive knowledge and metacognitive process: Important influences on mathematical performances. *Research & Teaching in Developmental Education, 2*(2), 34-39.

Garofalo, J. (1989). Beliefs and their influence on mathematical performance. *The Mathematics Teacher, 82*(7), 502-505. Retrieved September 8, 2020, from http://www.jstor.org/stable/27966379

Garofalo, J., & Lester, F. K. (1985). Metacognition, cognitive monitoring, and mathematical performance. *Journal for Research in Mathematics Education, 16*(3), 163-176. https://doi.org/10.2307/748391

Goodburn, R., Higgins, S., Parsons, S., Wall, K., & Wright J. (2005). *Learning to learn for life: Research and practical examples for the Foundation Stage and Key Stage 1*. Campaign for Learning. Network Educational Press Ltd.

Gorard, S. (2013). *Research design: Creating robust approaches for the social sciences*. Sage.

Gorard, S. (2014). A proposal for judging the trustworthiness of research findings. *Radical Statistics*, Issue 110, 47-59. https://www.radstats.org.uk/no110/Gorard110.pdf

Gorard, S. (2015). Rethinking 'quantitative' methods and the development of new researchers. *Review of Education, 3*(1), 72–96. https://doi.org/10.1002/rev3.3041

Grant, S., Mayo-Wilson, E., Montgomery, P., Michie, S., Hopewell, S., & Moher, D. (2018). CONSORT-SPI 2018 explanation and elaboration: Guidance for reporting social and psychological intervention trials. *Trials, 19*(1), 406. https://doi.org/10.1186/s13063-018-2735-z

Gross, T. (2022, April 28). How social-emotional learning became a target for Ron DeSantis and conservatives. NPR. https://www.npr.org/2022/04/28/1095042273/ron-desantis-florida-textbooks-social-emotional-learning

Hak, T., van Rhee, H., Suurmond, R. (2018). How to interpret results of meta-analysis. Erasmus Research Institute of Management. Rotterdam, The Netherlands. https://www.erim.eur.nl/fileadmin/erim_content/images/meta-essentials/How_to_interpret_results_of_meta-analysis_1.4.pdf

Harris, D., Lowrie, T., Logan, T., & Hegarty, M. (2021). Spatial reasoning, mathematics, and gender: Do spatial constructs differ in their contribution to performance? *British Journal of Educational Psychology, 91*(1), e12371. https://doi.org/10.1111/bjep.12371

Hart, C. (2001). *Doing a literature search: A comprehensive guide for the social sciences.* Sage.

Hattie, J. (2005). What is the nature of evidence that makes a difference to teaching? Australian Council for Educational Research (ACER) Conference 2005 – Using Data to Support Learning. http://research.acer.edu.au/research_conference_2005/7

Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research, 66*(2), 99–136. https://www.jstor.org/stable/1170605

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77* (1): 81–112. https://doi.org/10.3102/003465430298487

Hedges, L. V. (2012). Design of empirical research. In J. Arthur, M. Waring, R. Coe, & L. V. Hedges (Eds.), *Research Methods and Methodologies in Education* (pp. 23-30). Sage.

*Heemsoth, T. & Heinze, A. (2016). Secondary school students learning from reflections on the rationale behind self-made errors: A field experiment. *The Journal of Experimental Education, 84*(1), 98–118. https://doi.org/10.1080/00220973.2014.963215

Hendry, G. D. (1996). Constructivism and Educational Practice. *Australian Journal of Education*, *40*(1), 19–45. https://doi.org/10.1177/000494419604000103

Higgins, S. (2018). *Improving learning: Meta-analysis of intervention research in education.* Cambridge UP. https://doi.org/10.1017/9781139519618

Higgins, S., Baumfield, V., Lin, M., Moseley, D., Butterworth, M., Downey, G., Gregson, M., Oberski, I., Rockett, M., & Thacker, D. (2004). Thinking skills approaches to effective teaching and learning: What is the evidence for impact on learners. In *Research Evidence in Education Library* (p. 106). EPPI-Centre, Social Science Research Unit, Institute of Education. https://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/t_s_rv1.pdf?ver=2006-03-02-125124-377

Higgins, S., & Hall, E. (2004, September 16-18). Picking the strawberries out of the jam: Thinking critically about systematic reviews and meta-analysis. BERA 2004 Conference, Manchester, United Kingdom. http://www.leeds.ac.uk/educol/documents/00003835.doc

Higgins, S., Hall, E., Baumfield, V., & Moseley, D. (2005). A meta-analysis of the impact of the implementation of thinking skills approaches on pupils. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London. https://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/t_s_rv2.pdf?ver=2006-03-02-125128-393

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D. G., Barbour, V., Macdonald, H., Johnston, M., Lamb, S. E., Dixon-Woods, M., McCulloch, P., Wyatt, J. C., Chan, A.-W., & Michie, S. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *BMJ, 348* (g1687). https://doi.org/10.1136/bmj.g1687

Houk, S. (2000). *"Psychological care of infant and child": A reflection of its author and his times.* http://www.mathcs.duq.edu/~packer/DevPsych/Houk2000.html

*Hughes, E. M., Lee, J.-Y., Cook, M. J., & Riccomini, P. J. (2019). Exploratory study of a self-regulation mathematical writing strategy: Proof-of-concept. *Learning Disabilities: A Contemporary Journal, 19.* https://files.eric.ed.gov/fulltext/EJ1234949.pdf

IntHout, J., Ioannidis, J. P. A., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open, 6*(7), e010247. https://doi.org/10.1136/bmjopen-2015-010247

*Jackson Jackson, S. F. (2012). *Self regulated and communal learning contexts as they relate to math achievement and math self efficacy among African American elementary level students* [Doctoral dissertation, Howard University]. ProQuest Dissertations & Theses Global. http://www.proquest.com/pqdtglobal/docview/1435635242/abstract/806402A2C6F74460PQ/1

*Jacobse, A. E., & Harskamp, E. G. (2009). Student-controlled metacognitive training for solving word problems in primary school mathematics. *Educational Research and Evaluation, 15*(5), 447–463. https://doi.org/10.1080/13803610903444519

Jastrow, J. (1929). Review of The Ways of Behaviorism; Psychological Care of Infant and Child, John B. Watson; The Battle of Behaviorism, John B. Watson [Review of *Review of The Ways of Behaviorism; Psychological Care of Infant and Child, John B. Watson; The Battle of Behaviorism, John B. Watson*, by J. B. Watson & Wm. McDougall]. *Science*, *69*(1791), 455–457. http://www.jstor.org/stable/1653206

*Jitendra, A. K., Harwell, M. R., Dupuis, D. N., Karl, S. R., Lein, A. E., Simonson, G., & Slater, S. C. (2015). Effects of a research-based intervention to improve seventh-grade students' proportional problem solving: A cluster randomized trial. *Journal of Educational Psychology, 107*(4), 1019–1034. https://doi.org/10.1037/edu0000039

Kajander, A. E. (1999). Creating opportunities for children to think mathematically. *Teaching Children Mathematics, 5*(8), 480–486. http://www.jstor.org/stable/41197272

*Kang, Y. (2010). *Self-regulatory training for helping students with special needs to learn mathematics* [Doctoral dissertation, The University of Iowa]. ProQuest Dissertations & Theses Global. http://www.proquest.com/pqdtglobal/docview/756742069/abstract/3E7F85567FC3403FPQ/1

Kogut, A., Foster, M., Ramirez, D., & Xiao, D. (2019). Critical appraisal of mathematics education systematic review search methods: Implications for social sciences librarians. *College & Research Libraries, 80*(7). https://doi.org/10.5860/crl.80.7.973

Kotok, S. (2017). Unfulfilled Potential: High-Achieving Minority Students and the High School Achievement Gap in Math. *The High School Journal*, *100*(3), 183–202. https://www.jstor.org/stable/90024211

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. Educational Researcher, 49(4), 241–253. https://doi.org/10.3102/0013189X20912798

*Kramarski, B. & Dudai, V. (2009). Group-metacognitive support for online inquiry in mathematics with differential self-questioning. *Journal of Educational Computing Research, 40*(4), 377–404. https://doi.org/10.2190/EC.40.4.a

*Kramarski, B. & Friedman, S. (2014). Solicited versus unsolicited metacognitive prompts for fostering mathematical problem solving using multimedia. *Journal of Educational Computing Research, 50*(3), 285–314. https://doi.org/10.2190/EC.50.3.a

*Kramarski, B. & Gutman, M. (2006). How can self-regulated learning be supported in mathematical E-learning environments? *Journal of Computer Assisted Learning, 22*(1), 24–33. https://doi.org/10.1111/j.1365-2729.2006.00157.x

Kramarski, B., & Mevarech, Z. R. (2003). Enhancing mathematical reasoning in the classroom: The effects of cooperative learning and metacognitive training. *American Educational Research Journal*, *40*(1), 281–310. https://doi.org/10.3102/00028312040001281

*Kramarski, B., Weisse, I., & Kololshi-Minsker, I. (2010). How can self-regulated learning support the problem solving of third-grade students with mathematics anxiety? *ZDM, 42*(2), 179–193. https://doi.org/10.1007/s11858-009-0202-8

*Kramarski, B. & Zoldan, S. (2008). Using errors as springboards for enhancing mathematical reasoning with three metacognitive approaches. *The Journal of Educational Research, 102*(2), 137–151. https://doi.org/10.3200/JOER.102.2.137-151

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41(*4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Kuhfeld, M. & Lewis, K. (2022). *Student achievement in 2021–2022: Cause for hope and continued urgency.* NWEA Research. https://www.nwea.org/uploads/2022/07/Student-Achievement-in-2021-22-Cause-for-hope-and-concern.researchbrief-1.pdf

Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science, 9*(5), 178-181.

Lee, Y., Capraro, M. M., Capraro, R. M., & Bicer, A. (2018). A meta-analysis: improvement of students' algebraic reasoning through metacognitive training. *International Education Studies, 11*(10), 42–49. eric. https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1192530&site=ehost-live

*Lee, N. H., Yeo, D. J. S., & Hong, S. E. (2014). A metacognitive-based instruction for Primary Four students to approach non-routine mathematical word problems. *ZDM, 46*(3), 465–480. https://doi.org/10.1007/s11858-014-0599-6

*Lestari, W. & Jailani. (2018). Enhancing an ability mathematical reasoning through metacognitive strategies. *Journal of Physics: Conference Series, 1097*, 012117. https://doi.org/10.1088/1742-6596/1097/1/012117

Lewis, A. B. (1989). Training students to represent arithmetic word problems. *Journal of Educational Psychology, 81*(4), 521-531.

Lilford, R. J., & Jackson, J. (1995). Equipoise and the ethics of randomization. *Journal of the Royal Society of Medicine, 88*, 552-559.

Lin, L., & Aloe, A. M. (2021). Evaluation of various estimators for standardized mean difference in meta-analysis. *Statistics in Medicine, 40*(2), 403–426. https://doi.org/10.1002/sim.8781

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

Loeb, S., Bridges, M., Bassok, D., Fuller, B., & Rumberger, R. W. (2007). How much is too much? The influence of preschool centers on children's social and cognitive development. *Economics of Education Review, 26*(1), 52–66. https://doi.org/10.1016/j.econedurev.2005.11.005

Lopez-Garrido, G. (2021, January 7). *Structuralism and Titchener*. Simply Psychology. www.simplypsychology.org/structuralism.html

*Mandaci Şahin, S., & Kendir, F. (2013). The effect of using metacognitive strategies for solving geometry problems on students' achievement and attitude. *Educational Research and Reviews, 8*(19), 1777-1792. https://academicjournals.org/journal/ERR/article-full-text-pdf/EE6CB4C41375

Masters, G. (2018, August 27) The role of evidence in teaching and learning. *Teacher Magazine.* https://www.teachermagazine.com/au_en/articles/the-role-of-evidence-in-teaching-and-learning

Mayer, R. E. (1998). Cognitive, metacognitive, and motivational aspects of problem-solving. *Instructional Science, 26,*49-63.

*McClelland, M. M., Tominey, S. L., Schmitt, S. A., Hatfield, B. E., Purpura, D. J., Gonzales, C. R., & Tracy, A. N. (2019). Red Light, Purple Light! Results of an intervention to promote school readiness for children from low-income backgrounds. *Frontiers in Psychology, 10*, 2365. https://doi.org/10.3389/fpsyg.2019.02365

*Mevarech, Z. R., & Amrany, C. (2008). Immediate and delayed effects of meta-cognitive instruction on regulation of cognition and mathematics achievement. *Metacognition and Learning, 3*(2), 147–157. https://doi.org/10.1007/s11409-008-9023-3

Mevarech, Z. R., & Kramarski, B. (1997). IMPROVE: A multidimensional method for teaching mathematics in heterogeneous classrooms. *American Educational Research Journal, 34*(2), 365–394. https://doi.org/10.3102/00028312034002365

*Mevarech, Z. R., Terkieltaub, S., Vinberger, T., & Nevet, V. (2010). The effects of meta-cognitive instruction on third and sixth graders solving word problems. *ZDM, 42*(2), 195–203. https://doi.org/10.1007/s11858-010-0244-y

Meyer, J., & Land, R. (2003). Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practising within the disciplines. *ETL Project, Occasional Report 4.* http://www.etl.tla.ed.ac.uk/docs/ETLreport4.pdf

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ, 339*(b2535). https://doi.org/10.1136/bmj.b2535

*Morales, Z. A. (2016). *Using a repeated measures ANOVA design to analyze the effect Writing in mathematics has on the mathematics achievement of third grade English language learners and English speakers* [Doctoral dissertation, Florida International University]. ProQuest Dissertations & Theses Global. https://www.proquest.com/docview/2014472693/abstract/8D27B751F8B54B72PQ/1

Morrison, K. (2021). *Taming randomised controlled trials in education: Exploring key claims, issues, and debates.* Routledge: Taylor & Francis Group. https://doi.org/10.4324/9781003042112

Moshman, D. (2018). Metacognitive theories revisited. *Educational Psychology Review, 30*(2), 599–606. https://doi.org/10.1007/s10648-017-9413-7

*Motteram, G., Choudry, S., Kalambouka, A., Barton, A., Hutcheson, G., Onat-Stelma, Z., & Bragg, J. (2016, November 4). ReflectED metacognition evaluation report: An approach to improving learning skills using digital technology. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/our-work/projects/reflected-meta-cognition/

National Library of Medicine. (2021, December 27). *Introduction to MeSH.* https://www.nlm.nih.gov/mesh/introduction.html

Norman, E. (2020). Why Metacognition Is Not Always Helpful. *Frontiers in Psychology*, *11*. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01537

Norman, G. (2003). RCT = results confounded and trivial: The perils of grand educational experiments. *Medical Education, 37*(7), 582–584.

*O'Neal, L. (2015). *The effects of metacognitive writing on student achievement in Advanced Placement® calculus* [Doctoral dissertation, Seattle Pacific University]. Education Dissertations. https://digitalcommons.spu.edu/soe_etd/3

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, *8*(2), 157–159. https://doi.org/10.3102/10769986008002157

*Ozsoy, G., & Ataman, A. (2009). The effect of metacognitive strategy training on mathematical problem solving achievement. *International Electronic Journal of Elementary Education, 1*(2), 68–83. https://www.iejee.com/index.php/IEJEE/article/view/278

Palinscar, A. S. & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1(*2), 117-175. https://doi.org/10.1207/s1532690xci0102_1

Panaoura, A. (2012). Improving problem solving ability in mathematics by using a mathematical model: A computerized approach. *Computers in Human Behavior, 28*(6), 2291–2297. https://doi.org/10.1016/j.chb.2012.06.036

Papert, S. (1999, March 29). Child psychologist Jean Piaget. *Time*. http://content.time.com/time/subscriber/article/0,33009,990617-2,00.html

*Pappas Schattman, S. S. (2005). *Fostering kindergarteners' metacognition* [Doctoral dissertation, Columbia University]. ProQuest Dissertations & Theses Global. http://www.proquest.com/docview/305012525/abstract/BF17374CC086452BPQ/1

*Pennequin, V., Sorel, O., Nanty, I., & Fontaine, R. (2010). Metacognition and low achievement in mathematics: The effect of training in the use of metacognitive skills to solve mathematical word problems. *Thinking & Reasoning, 16*(3), 198–220. https://doi.org/10.1080/13546783.2010.509052

*Perels, F., Dignath, C., & Schmitz, B. (2009). Is it possible to improve mathematical achievement by means of self-regulation strategies? Evaluation of an intervention in regular math classes. *European Journal of Psychology of Education, 24*(1), 17. https://doi.org/10.1007/BF03173472

Perkins, D. N. & Salomon, G. (1989) Are cognitive skills context-bound? *Educational Researcher, 18*(1), 16-25. https://doi.org/10.3102/0013189X018001016

Perry, J., Lundie, D., & Golder, G. (2019). Metacognition in schools: What does the literature suggest about the effectiveness of teaching metacognition in schools? *Educational Review, 71*(4), 483–500. https://doi.org/10.1080/00131911.2018.144112

Peterson, J., Pearce, P. F., Ferguson, L. A., & Langford, C. A. (2017). Understanding scoping reviews: Definition, purpose, and process. J*ournal of the American Association of Nurse Practitioners, 29*(1), 12–16. https://doi.org/10.1002/2327-6924.12380

Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research, 31*(6), 459-470. https://doi.org/10.1016/S0883-0355(99)00015-4.

Pintrich, P. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). San Diego: Academic Press.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory Into Practice, 41*(4), 219-225. https://doi.org/10.1207/s15430421tip4104_3

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33. https://doi.org/10.1037/0022-0663.82.1.33

Pólya, G. (1954). *Induction and analogy in mathematics: Volume I of mathematics and plausible reasoning.* Princeton University Press.

Pólya, G. (1971). *How to solve it* (2nd ed.). Princeton University Press.

Quigley, A., Muijs, D., & Stringer, E. (2018). Metacognition and self-regulated learning: Guidance report. London: Education Endowment Foundation. https://d2tic4wvo1iusb.cloudfront.net/eef-guidance reports/metacognition/EEF_Metacognition_and_self-regulated_learning.pdf?v=1678650171

Reeve, R. A., & Brown, A. L. (1985). Metacognition reconsidered: Implications for intervention research. *Journal of Abnormal Child Psychology, 13*(3), 343–356. https://doi.org/10.1007/BF00912721

Rellinger, E., Borkowski, J. G., Turner, L. A., & Hale, C. A. (1995). Perceived task difficulty and intelligence: Determinants of strategy use and recall. Intelligence, 20(2), 125–143. https://doi.org/10.1016/0160-2896(95)90029-2

*Riggs, R. M. (2012). *Can practice calibrating by test topic improve public school students' calibration accuracy and performance on tests?* [Doctoral dissertation, Old Dominion University]. ProQuest Dissertations & Theses Global. http://www.proquest.com/docview/1138333949/abstract/44B7EBCE65C64956PQ/1

*Rizk, N., Attia, K., & Al-Jundi, A. (2017). The impact of metacognition strategies in teaching mathematics among innovative thinking students in primary school, Rafha, KSA. *International Journal of English Linguistics, 7*(3), 103. https://doi.org/10.5539/ijel.v7n3p103

Roediger, H. L. (2004). What happened to behaviorism. *APS Observer*, *17*(3). https://www.psychologicalscience.org/observer/what-happened-to-behaviorism

Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641.

Roser, M., & Ortiz-Ospina, E. (2016). *Literacy.* Our World in Data. https://ourworldindata.org/literacy

*Sarette, S. A. (2014). *The building blocks of a brain training community: How early experiences in a second grade classroom shape the development of self-regulation and working memory skills of children* [Doctoral dissertation, New England College]. ProQuest Dissertations & Theses Global. http://www.proquest.com/docview/1616644921/abstract/F875B361847948D3PQ/1

Schleicher, A. (n.d.). *Skills matter: Additional results from the survey of adult skills* [Slideshow]. OECD. https://www.oecd.org/skills/piaac/

Schleicher, A. (2019). *PISA 2018: Insights and interpretations.* OECD.org. https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf

*Schmitt, S. A. (2013). *Strengthening school readiness for children at risk: Evaluating self-regulation measures and an intervention using classroom games* [Doctoral dissertation, Oregon State University]. ProQuest Dissertations & Theses Global. https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED555778&site=ehost-live

*Schmitt, S. A., McClelland, M. M., Tominey, S. L., & Acock, A. C. (2015). Strengthening school readiness for Head Start children: Evaluation of a self-regulation intervention. *Early Childhood Research Quarterly, 30*, 20–31. https://doi.org/10.1016/j.ecresq.2014.08.001

Schoenfeld, A. H. (1987). What's all the fuss about metacognition? In A. H. Schoenfeld (Ed.), *Cognitive science and mathematics education* (p. 189-216). Lawrence Erlbaum.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (p. 334–370). Macmillan Publishing Co, Inc.

Schoenfeld, A. H. (2020). Mathematical practices, in theory and practice. *ZDM, 52*(6), 1163–1175. https://doi.org/10.1007/s11858-020-01162-w

Schraw, G. & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*(4), 351-371. https://doi.org/10.1007/BF02212307

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist, 26*(3-4), 207–231. https://doi.org/10.1207/s15326985ep2603&4_2

See, B. H., Gorard, S., & Siddiqui, N. (2016). Teachers' use of research evidence in practice: A pilot study of feedback to enhance learning. *Educational Research, 58*(1), 56–72. https://doi.org/10.1080/00131881.2015.1117798

Seegers, G., & Boekaerts, M. (1996). Gender-related differences in self-referenced cognitions in relation to mathematics. *Journal for Research in Mathematics Education, 27*(2), 215–240. https://doi.org/10.2307/749601

*Shamir, A., & Lifshitz, I. (2013). E-Books for supporting the emergent literacy and emergent math of children at risk for learning disabilities: Can metacognitive guidance make a difference? *European Journal of Special Needs Education, 28*(1), 33–48. https://doi.org/10.1080/08856257.2012.742746

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ, 349*(g7647). https://doi.org/10.1136/bmj.g7647

*Shilo, A., & Kramarski, B. (2019). Mathematical-metacognitive discourse: How can it be developed among teachers and their students? Empirical evidence from a videotaped lesson and two case studies. *ZDM, 51*(4), 625–640. https://doi.org/10.1007/s11858-018-01016-6

Siegler, R. S. (1994). Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Science, 3*(1), 1–5. https://doi.org/10.1111/1467-8721.ep10769817

Siegler, R. S., & Chen, Z. (1998). Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology, 36,* 273–310. https://doi.org/10.1006/cogp.1998.0686

*Sings Jenkins, J. (2009). *The effects of explicit self-regulated learning strategy instruction on mathematics achievement* [Doctoral dissertation, University of North Carolina]. ProQuest Dissertations & Theses Global. https://www.proquest.com/openview/5b6082c89190396c3976eecc2b28c881/1?pq-origsite=gscholar&cbl=18750

Slavin, R. E. (2008). Perspectives on evidence-based research in education: What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5-14. https://doi.org/10.3102/0013189X08314117

Slavin, R. E. (2013). Effective programmes in reading and mathematics: Lessons from the Best Evidence Encyclopedia. *School Effectiveness and School Improvement, 24*(4), 383-391. https://doi.org/10.1080/09243453.2013.797913

Slavin, R., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness, 4*(4), 370–380. https://doi.org/10.1080/19345747.2011.558986

Song, F., Hooper, L., & Loke, Y. K. (2013). Publication bias: What is it? How do we measure it? How do we avoid it? *Open Access Journal of Clinical Trials, 5*, 71–81. https://doi.org/10.2147/OAJCT.S34419

Sun-Lin, H.-Z., & Chiou, G.-F. (2017). Effects of self-explanation and game-reward on sixth graders' algebra variable learning. *Journal of Educational Technology & Society, 20*(4), 126–137. https://www.jstor.org/stable/26229211

Suurmond, R., van Rhee, H., & Hak, T. (2017). Introduction, comparison, and validation of Meta-Essentials: A free and simple tool for meta-analysis. *Research Synthesis Methods, 8*(4), 537–553. https://doi.org/10.1002/jrsm.1260

Tao, K., Li, X., Zhou, Q., Moher, D., Ling, C., & Yu, W. (2011). From QUOROM to PRISMA: A survey of high-impact medical journals' instructions to authors and a review of systematic reviews in anesthesia literature. PLoS ONE, 6(11), e27611. https://doi.org/10.1371/journal.pone.0027611

Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., & Bond, M. (2020). *EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis.* EPPI-Centre Software. London: UCL Social Research Institute. http://eppi.ioe.ac.uk/eppireviewer4/

*Tok, S. (2013). Effects of the Know-Want-Learn strategy on students' mathematics achievement, anxiety and metacognitive skills. *Metacognition and Learning, 8*(2), 193–212. https://doi.org/10.1007/s11409-013-9101-z

*Tominey, S. L., & McClelland, M. M. (2011). Red Light, Purple Light: Findings from a randomized trial using circle time games to improve behavioral self-regulation in preschool. *Early Education and Development, 22*(3), 489–519. https://doi.org/10.1080/10409289.2011.574258

Tomic, W. (1993). Behaviorism and cognitivism in education. *Psychology: A Journal of human behavior, 30*(3/4), 38-46. https://research.ou.nl/ws/files/9520563/BEHAVIORISM%20AND%20COGNITIVISM%20IN%20EDUCATION.pdf

Torgerson, C. (2003). *Systematic reviews*. Continuum.

Torgerson, C., Hall, J., & Light, K. (2012). Systematic reviews. In J. Arthur, M. Waring, R. Coe, & L. V. Hedges (Eds.), *Research methods & methodologies in education* (pp. 217-230). Sage.

*Tzohar-Rozen, M., & Kramarski, B. (2013). How does an affective self-regulation program promote mathematical literacy in young students? *Hellenic Journal of Psychology, 10*, 211-234. https://pseve.org/wp-content/uploads/2018/03/Volume10_Issue3_Tzohar-Rozen.pdf

*Tzohar-Rozen, M., & Kramarski, B. (2017). Metacognition and meta-affect in young students: Does it make a difference in mathematical problem solving? *Teachers College Record, 119*(13). https://doi.org/10.1177/016146811711901308

*Ubuz, B., & Erdoğan, B. (2019). Effects of physical manipulative instructions with or without explicit metacognitive questions on geometrical knowledge acquisition. *International Journal of Science and Mathematics Education, 17*(1), 129–151. https://doi.org/10.1007/s10763-017-9852-0

Verschaffel, L., Depaepe, F., & Mevarech, Z. (2019). Learning mathematics in metacognitively oriented ICT-based learning environments: A systematic review of the literature. *Education Research International, 2019*, 1–19. https://doi.org/10.1155/2019/3402035

The Visible Learning Research. (n.d.). Corwin Visible Learning Plus. https://www.visiblelearning.com/content/visible-learning-research

Von Glasersfeld, E. (1982). An interpretation of Piaget's constructivism. *Revue Internationale de Philosophie, 36*(142/143 (4)), 612–635. https://www.jstor.org/stable/23945415

*Vula, E., Avdyli, R., Berisha, V., Saqipi, B., & Elezi, S. (2017). The impact of metacognitive strategies and self-regulating processes of solving math word problems. *International Electronic Journal of Elementary Education, 10*(1), 49–59. https://www.iejee.com/index.php/IEJEE/article/view/298

*Wang, A. Y., Fuchs, L. S., Fuchs, D., Gilbert, J. K., Krowka, S., & Abramson, R. (2019). Embedding self-regulation instruction within fractions intervention for third graders with mathematics difficulties. *Journal of Learning Disabilities*, *52*(4), 337–348. https://doi.org/10.1177/0022219419851750

Wang, Y., & Sperling, R. A. (2020). Characteristics of effective self-regulated learning interventions in mathematics classrooms: A systematic review. Frontiers in Education, 5. https://doi.org/10.3389/feduc.2020.00058

Waters, H. S., & Kunnmann, T. W. (2010). Metacognition and strategy discovery in early childhood. In H. S. Waters & W. Schneider (Eds.), *Metacognition, strategy use, and instruction* (pp. 3–22). The Guilford Press.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*(2), 158–177. https://doi.org/10.1037/h0074428

Watson, J. B., & Watson, R. R. (1928). *Psychological care of infant and child.* W. W. Norton & Co. (Reprint edition 1972, Arno Press Inc.)

Weiss, J. (2005, July 22). *Back to basics through the years*. The Chicago Reporter. http://www.chicagoreporter.com/back-basics-through-years/

Wessman Huber, C. (2010). *The Impact of reciprocal teaching on mathematics problem solving for grade 4 students* [Doctoral dissertation, Central Connecticut State University]. ProQuest Dissertations & Theses Global. www.proquest.com/dissertations-theses/impact-reciprocal-teaching-on-mathematics-problem/docview/914241593/se-2

White, M. B., & Hall, A. E. (1980). An overview of intelligence testing. *Educational Horizons*, *58*(4), 210–216. https://www.jstor.org/stable/42924403

*Wijaya, A., Heuvel-Panhuizen, M. V. den, Doorman, M., & Veldhuis, M. (2018). Opportunity-to-learn to solve context-based mathematics tasks and students' performance in solving these tasks – Lessons from Indonesia. *Eurasia Journal of Mathematics, Science and Technology Education, 14*(10), em1598. https://doi.org/10.29333/ejmste/93420

Wilson, D. B. (n.d.) *Practical meta-analysis effect size calculator.* Campbell Collaboration. https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.phpf

Winne, P. H. (1995). Inherent details in self-regulated learning. *Educational Psychologist, 30*(4), 173–187. https://doi.org/10.1207/s15326985ep3004_2

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences, 8*(4), 327–353. https://doi.org/10.1016/S1041-6080(96)90022-9

Wright, B. L. (2022, August 31). *Pre-pandemic, more U.S. students were excelling in math.* The Thomas B. Fordham Institute. https://fordhaminstitute.org/national/commentary/pre-pandemic-more-us-students-were-excelling-math

Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology*, *11*(4), 307–313. https://doi.org/10.1016/0361-476X(86)90027-5

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25(*1), 3-17. https://doi.org/ 10.1207/s15326985ep2501_2

Zimmerman, B. J. (1995). Self-regulation involves more than metacognition: A social cognitive perspective. *Educational Psychologist, 30*(4), 217–221. https://doi.org/10.1207/s15326985ep3004_8

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice, 41*, 64–70. https://doi.org/10.1207/s15430421tip4102_2