Case Study 1: An Evidence-Based Practice Review Report

Theme: School/setting Based Interventions for Learning.

How effective is metacognition instruction at improving the word problem-solving of children who are low-achievers in maths?

Summary

Metacognitive knowledge and skills are powerful predictors of academic outcomes (Wang, Haertel, & Walberg, 1990) but are often lacking in children who are low-achievers in maths (Miller & Mercer, 1997). Metacognition instruction seeks to address this deficit (Veenman, 2015). Interventions address strategy knowledge (plan-monitor-evaluate), task knowledge (when and why to apply strategies), and person knowledge (strengths, weaknesses, and motivation), and provide opportunity to practise (Flavell, 1979; Livingston, 1996; Pintrich, 2002). This review sought to evaluate the effect of metacognition interventions on mathematical word problem-solving. A systematic literature search was conducted, identifying seven studies for review. A meta-analysis showed a large combined effect size when comparing intervention to comparison participants. This, in combination with sufficient methodological quality among the reviewed studies, suggests metacognition instruction can be recommended as evidence-based practice (Gersten, Fuchs, Compton, Coyne, Greenwood, & Innocenti, 2005). Recommendations for educational psychology practice, limitations of the review, and recommendations for future research are discussed.

Introduction

Metacognition

Metacognition refers to knowledge about, and regulation of, cognition (Schraw, 1998). While cognitive skills are used to perform tasks (such as multiplication), metacognitive skills are used to decide how to perform tasks and to evaluate performance (Garner, 1987). Psychological research on metacognition began in earnest in the 1970s (Gleitman et al., 1972). It has since been acknowledged as a concept of profound psychological importance, being incorporated into a revision of Bloom's Taxonomy of Learning as a fourth dimension of knowledge (Krathwohl, 2002).

There are two prominent theoretical models of metacognition. Flavell (1979) distinguishes four components: knowledge, experience, goals, and actions (see Figure 1). *Knowledge* comprises three sub-components: person, task, and strategy. 'Person knowledge' involves awareness of oneself and others as cognitive processors, including strengths, weaknesses, and motivation. 'Task knowledge' involves awareness of how to manage cognitive enterprises, including implications of task difficulty and situational norms for strategy selection (Pintrich, 2002). 'Strategy knowledge' involves awareness of ways of effectively achieving cognitive goals, including planning, monitoring, evaluating, information-acquisition strategies (e.g. mnemonics), and problem-solving heuristics. Metacognitive *experiences* are conscious feelings accompanying cognitive enterprises, such as being aware that one doesn't understand something. *Goals* refer to awareness of task objectives and *actions* refer to strategies or behaviours employed to achieve goals.

An alternative model (Schraw, 1998) distinguishes two components: knowledge and regulation (see Figure 2). *Knowledge* comprises three sub-components: declarative, procedural, and conditional. 'Declarative knowledge' involves awareness about oneself and factors influencing performance. 'Procedural knowledge' involves awareness of effective strategies and heuristics to complete tasks. 'Conditional knowledge' involves awareness of

when and why to use declarative and procedural knowledge, such as allocating resources and selecting strategies. *Regulation* is the active employment of knowledge before, during, and after a task to plan, monitor, and evaluate learning and performance.

Conceptual similarities across the models include self-awareness of strengths, weaknesses, and motivation; knowledge of how and when to use cognitive strategies; and the planning, monitoring and evaluating sequence. Psychometric evidence supports the parsimony of a two-factor model (Schraw & Dennison, 1994). Unrestricted factor analysis of a 52-item metacognitive inventory produced an unreliable six-factor solution but this did not map onto the six conceptual sub-components. Restricted factor analysis, however, strongly supported a two-factor solution (knowledge and regulation), with high internal consistency ($\alpha = .91$) on each factor and 44 items loading unambiguously onto a single factor. Furthermore, the factors contributed separately to performance on a reading comprehension test, suggesting the need to develop both metacognitive knowledge and regulation skills for optimal outcomes.

Figure 1

A Four-Component Model of Metacognition, Adapted From Flavell (1979)

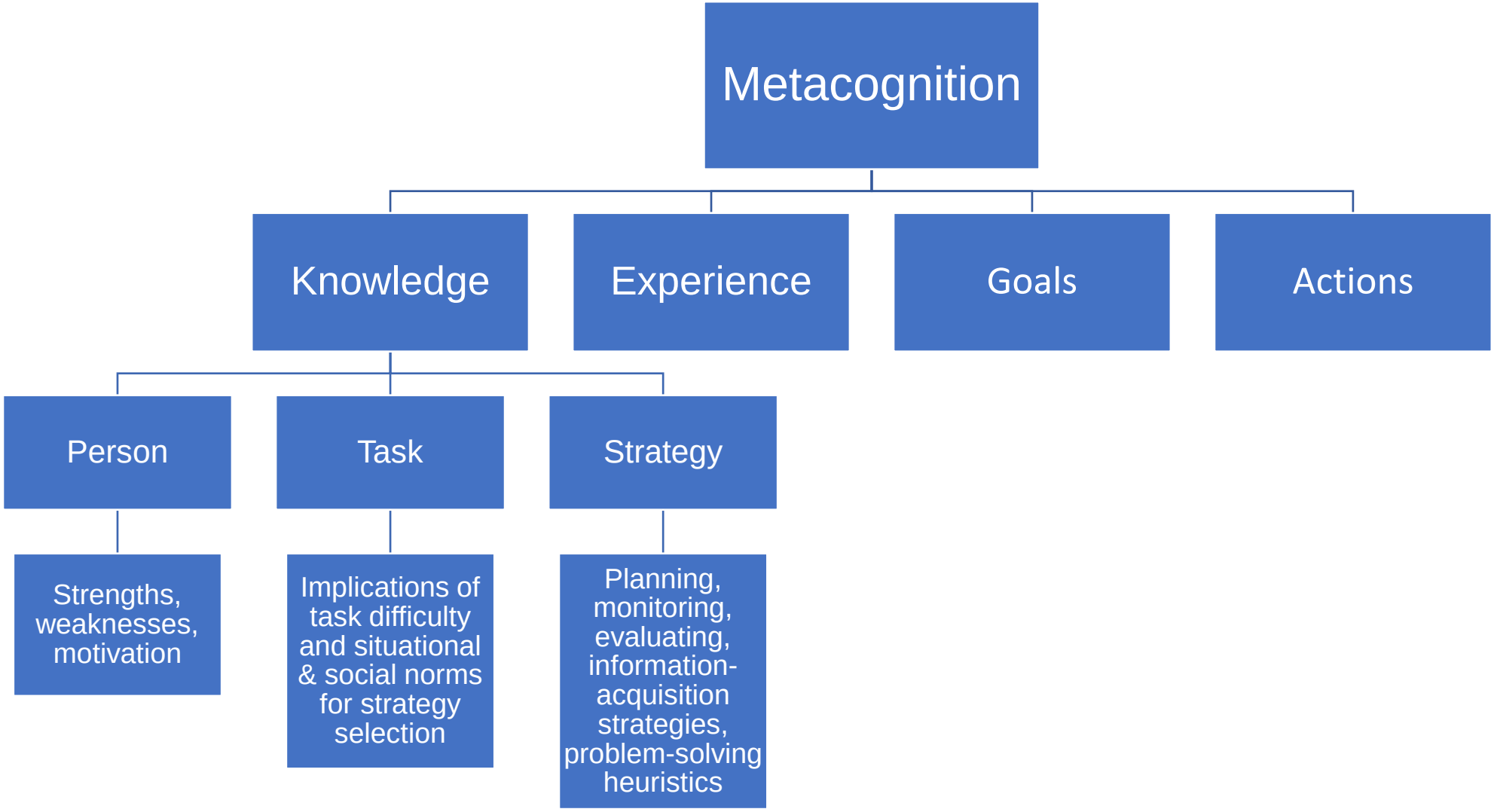
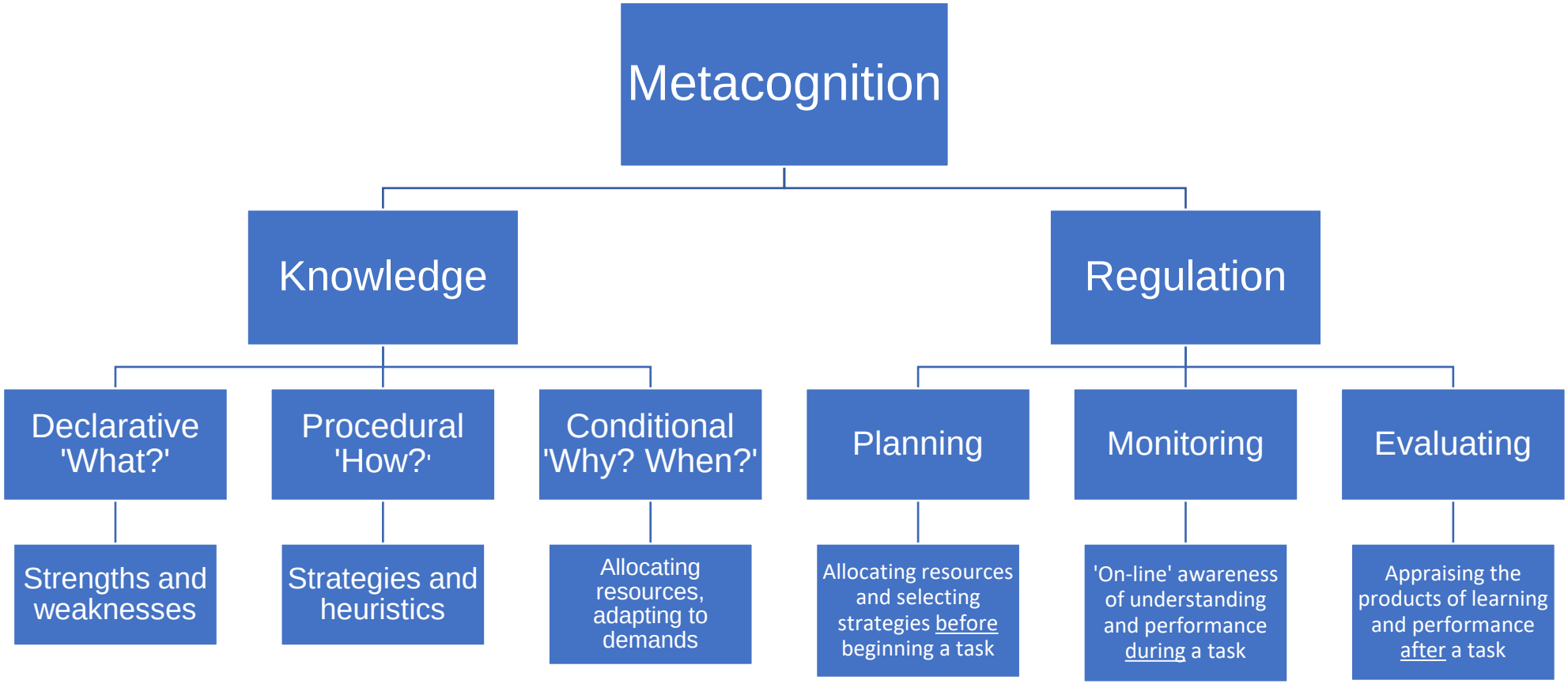


Figure 2

A Two-Component Model of Metacognition, Adapted From Schraw (1998)



Teaching Metacognition

According to the Education Endowment Foundation, teaching metacognition in schools has high impact for very low cost, based on extensive evidence (Quigley et al., 2018). Metacognition was identified as the single most important predictor of learning outcomes, above classroom management, student/teacher interactions and 27 further variables (Wang, Haertel, & Walberg, 1990).

Metacognitive skills may be acquired as part of typical development, emerging around age six and increasing in frequency and quality from age eight (Veenman & Spaans, 2005). Metacognition develops first as a domain-specific attribute but, with growing proficiency around age 14, generalises and promotes learning transfer between contexts (Schraw, 1998; Veenman & Spaans, 2005). Metacognitive skills developed in one subject-area benefit individuals in other subject-areas and in life beyond formal education (Pintrich, 2002).

However, some children require explicit instruction to acquire metacognitive skills (Veenman, 2015). An important area of learning that requires metacognition and has everyday relevance is mathematical problem-solving (Montague, 1997). This is a foundational skill for school attainment and is vital for everyday tasks such as grocery shopping. Most students struggle with metacognitive components such as assessing their ability, selecting appropriate strategies, organising information, monitoring, and evaluating outcomes (De Corte et al., 2000; Miller & Mercer, 1997).

A prominent intervention addressing metacognitive skills in mathematical problem-solving is Cognitive Strategy Instruction (CSI) (Montague et al., 2011). CSI combines metacognitive knowledge, regulation, and cognitive strategies in a sequential problem-solving model (see Table 1). Children memorise the seven ‘cognitive’ strategies and perform the ‘metacognitive’ strategies at each step to ensure they’ve completed the step comprehensively. For this author, CSI’s distinction between ‘cognitive’ and ‘metacognitive’ strategies is conceptually unclear since ‘hypothesise’ and ‘check’ are metacognitive in nature.

A systematic literature review of CSI identified five single-subject and two group-experimental designs (Montague & Dietz, 2009). Despite consistent evidence of the effectiveness of CSI at improving problem-solving, the review concluded that findings did not meet the methodological criteria for evidence-based practice. It suggested future research needed more stringent experimental designs. In light of this, the current review broadened the scope of intervention, seeking any study employing metacognitive instruction, but narrowed the scope of experimental design, seeking only studies with pre-/post-data on intervention and comparison groups. This is the most appropriate research design for answering questions about the effectiveness of interventions on outcomes (Petticrew & Roberts, 2003).

Table 1

Components of Cognitive Strategy Instruction, adapted from Montague et al. (2011)

Cognitive strategies	Metacognitive strategies
1) Read (for understanding)	1) Say
2) Paraphrase (your own words)	2) Ask
3) Visualise (a picture or diagram)	3) Check
4) Hypothesise (a plan to solve the problem)	
5) Estimate (predict the answer)	
6) Compute (do the arithmetic)	
7) Check (make sure everything is right)	

Review Question

This review aims to answer the question:

How effective is metacognition instruction at improving the word problem-solving of children who are low-achievers in maths?

This question is of relevance to educational psychologists because it will indicate the potential benefit of providing metacognition training to school staff and whether targeted metacognition interventions for low-achievers in mathematics are worth recommending to schools.

Critical Review of the Evidence Base

Literature Search

A systematic literature search was conducted on 15th December 2019 using three online databases: Web of Science, Educational Resources Information Center (ERIC), and PsycINFO. Search terms are listed in Table 2. Search results were limited to peer-reviewed journal articles. Ancestral and citation searches were conducted on articles selected for inclusion in the review.

Table 2

Terms Used in the Literature Searches

Search terms	Rationale
metacogniti* OR self-regulat* OR “cognitive strategy instruction” OR “cognitive-based instruction”	• This review seeks to evaluate the effectiveness of metacognition instruction • ‘Self-regulation’ or ‘self-regulatory’ were found through pilot searches to be common ways of referring to metacognition instruction • ‘Cognitive strategy instruction’ and ‘Cognitive-based instruction’ were found through pilot searches to be specific intervention approaches that included metacognitive instruction
dyscalculi* OR "number fact disorder" OR psychological difficulties in math* OR math* learning difficult* OR arithmetic learning difficult* OR arithmetic learning disabilit* OR math* learning disabilit* OR "low achiev*" in math*	• This review seeks to evaluate the effectiveness of metacognition instruction for CYP who are low-achievers or have identified learning difficulties in maths

Note. Truncation (*) was used to include any ending of root words. Speech marks (“”) were used to include exact phrase matching.

Article Screening

Database searches yielded 248 results. Following removal of 31 duplicates, 217 articles underwent title and abstract screening to determine eligibility for inclusion in the review (see Table 3 for criteria). 177 articles were excluded, leaving 40 articles for full text screening. 10 additional articles were identified through ancestral and citation searching and screened at full text. 43 studies were excluded (see Appendix A for reasons), leaving seven studies eligible for review (see Table 4). A flow diagram of article selection is provided in Figure 3.

Table 3

Criteria for Inclusion in the Review with Rationale

	Criterion	Inclusion	Exclusion	Rationale
1	Type of publication	The article is published in a peer-reviewed journal	The article is not published in a peer-reviewed journal	This ensures the article has been subject to quality control by trained researchers
2	Language of publication	The article is written in English (although the research can be conducted in any country)	The article is not written in English	This ensures the article can be understood by the author without being translated
3	Date of publication	The article is published on or before 15/12/2019	The article is published after 15/12/2019	This ensures all relevant articles available on the date of the literature search are included
4	Primary data	The article consists of original research	The article is a review or meta-analysis	This review seeks to evaluate original research
5	Intervention	At least one, but not all, groups	No group, or all groups, receive	This review seeks to evaluate the

	Criterion	Inclusion	Exclusion	Rationale
		receive metacognition instruction [a]	metacognition instruction	effect of including metacognition instruction
6	Experimental design	There is at least one intervention and one comparison group	The research does not use an intervention vs comparison design	This review seeks to compare the impact of metacognition instruction with other teaching methods
7	Outcome measures	At least one outcome measure is of mathematical problem-solving [b]	There is no outcome measure of mathematical problem-solving	This review seeks to evaluate the effects of metacognition instruction on problem-solving
8	Outcome data	There is quantitative data on problem-solving both pre- and post-intervention	There is only qualitative data or missing quantitative data	This review seeks to conduct a quantitative analysis
9	Setting	The research is conducted in a school or other purpose-built educational setting	The research is not conducted in a purpose-built educational setting e.g. home, hospital, experimental laboratory	This review seeks to evaluate school-/setting-based interventions
10	Age of participants	Participants are of English compulsory school age (5-16)	Participants are younger than 5 or older than 16	This review seeks to evaluate the effectiveness of metacognitive instruction for school-aged children

	Criterion	Inclusion	Exclusion	Rationale
11	Special educational needs (SEN) of participants	Participants have been identified either as having mathematical learning difficulties or as low-achievers in maths [c]	Participants have been identified only with other forms of SEN or have not been identified with SEN or data is not provided	This review seeks to evaluate the effectiveness of metacognitive instruction for children who are low-achievers in maths
12	Intervention delivery	The intervention is temporary	The intervention replaces regular teaching with the intention of being implemented permanently	This review seeks to evaluate the effectiveness of temporary interventions that could be recommended to schools by educational psychologists

[a] The nature and amount of metacognition instruction is evaluated in Weight of Evidence C; it was sufficient for a study to explicitly claim its intervention involved a metacognitive component for it to meet this inclusion criterion.

[b] Mathematical word problems involve calculations within a narrative frame.

[c] The extent of mathematical learning difficulties is not specified in the inclusion criteria due to the wide variability in definitions exhibited in the literature.

Figure 3

Flow Diagram of Literature Search and Article Screening

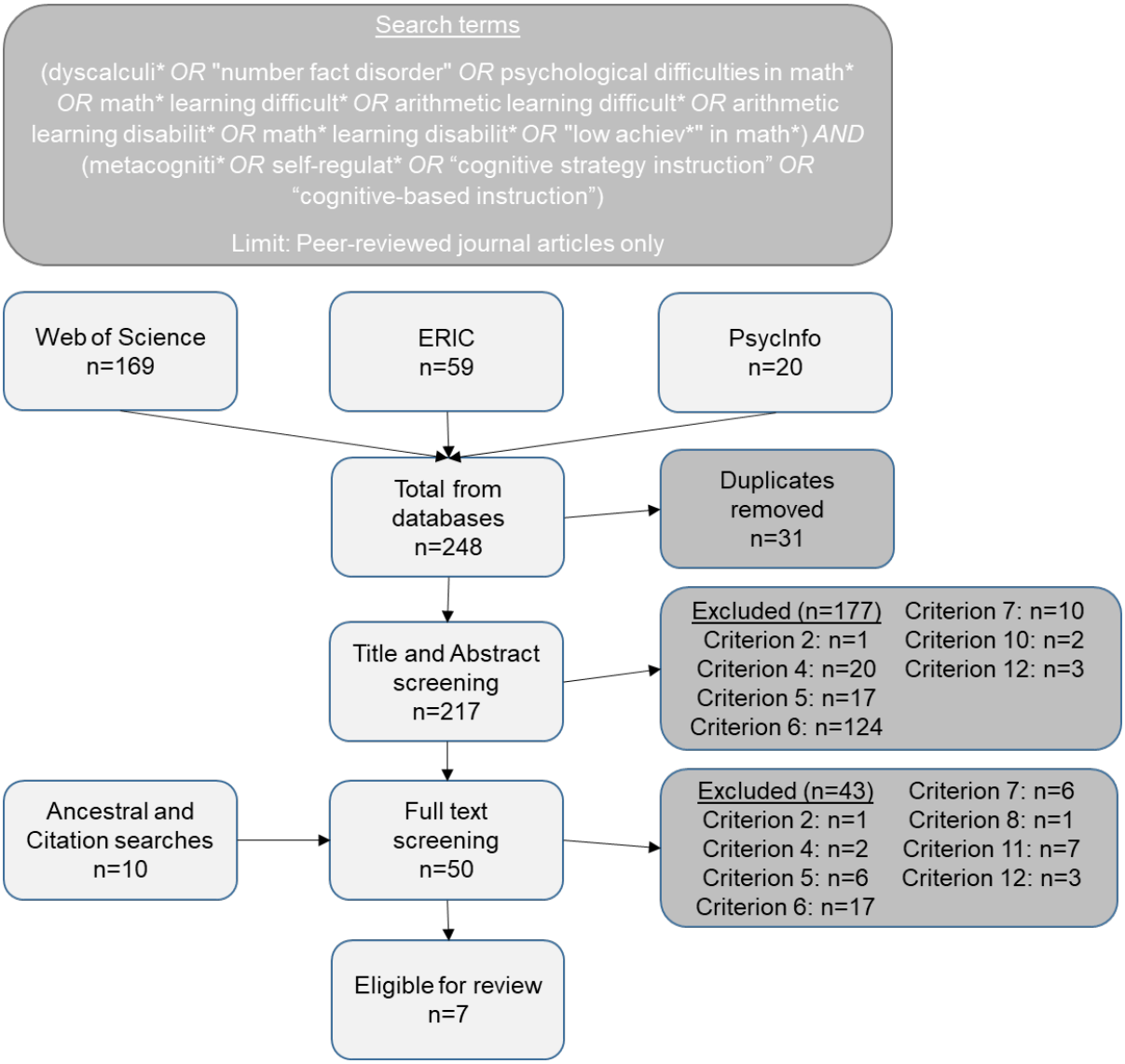


Table 4

References of Articles Included in the Review

	Reference of eligible article
1	Chung, K. K. H., & Tam, Y. H. (2005). Effects of cognitive-based instruction on mathematical problem solving by learners with mild intellectual disabilities. *Journal of Intellectual and Developmental Disability*, *30*(4), 207–216. https://doi.org/10.1080/13668250500349409
2	Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., & Schroeter, K. (2003). Enhancing third-grade students' mathematical problem solving with self-regulated learning strategies. *Journal of Educational Psychology*, *95*(2), 306–315. https://doi.org/10.1037/0022-0663.95.2.306
3	Kajamies, A., Vauras, M., & Kinnunen, R. (2010). Instructing low-achievers in mathematical word problem solving. *Scandinavian Journal Of Educational Research*, *54*(4), 335–355. https://doi.org/10.1080/00313831.2010.493341
4	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
5	Teong, S. K. (2003). The effect of metacognitive training on mathematical word-problem solving. *Journal of Computer Assisted Learning*, *19*(1), 46–55. https://doi.org/10.1046/j.0266-4909.2003.00005.x
6	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
7	Zhu, N. (2015). Cognitive strategy instruction for mathematical word problem-solving of students with mathematics disabilities in China. *International Journal of Disability, Development and Education*, *62*(6), 608–627. https://doi.org/10.1080/1034912X.2015.1077935

Weight of Evidence

The seven included studies were critically appraised using the Weight of Evidence (WoE) framework (Gough, 2007). Dimensions considered were methodological quality (WoE A), methodological relevance (WoE B), and topic relevance (WoE C).

WoE A was a generic judgment of the quality of the research design including participant description, intervention implementation, outcome measures, and data analysis. A published coding protocol was used to assess WoE A (Gersten et al., 2005). WoE B and C were judgments relating to the review question, using coding protocols developed by the author. WoE D is the average of WoE A, B, and C. A summary of WoE ratings is provided in Table 5. Full details of appraisal criteria are provided in Appendix B. Coding protocols for WoE A are provided in Appendix D.

Table 5

Summary of Weight of Evidence Ratings

Study	WoE A: Methodological quality	WoE B: Methodological relevance	WoE C: Topic relevance	WoE D: Overall rating
Chung et al. (2005)	1 Low	2 Medium	1.8 Medium	1.6 Medium
Fuchs et al. (2003)	3 High	2.5 High	2 Medium	2.5 High
Kajamies et al. (2010)	1 Low	1.75 Medium	2.4 Medium	1.72 Medium
Pennequin et al. (2015)	2 Medium	2.5 High	1.2 Low	1.9 Medium
Teong (2003)	0 Low	2.25 Medium	1.2 Low	1.15 Low
Wang et al. (2019)	3 High	2.5 High	1.8 Medium	2.43 Medium-High
Zhu (2015)	3 High	2.5 High	1.6 Medium	2.37 Medium-High

Note. WoE ratings are described as 'High' for scores ≥ 2.5, 'Medium' for scores ≥ 1.5 and < 2.5, and 'Low' for scores < 1.5.

Mapping the Field

The seven included studies used intervention vs comparison designs to explore the effect of metacognition instruction on word problem-solving of children who were low achievers in maths. Details of participants and procedures are provided in Table 6.

Table 6

Key Information about Participants and Procedures of the Seven Included Studies

Study	Participants	Procedure
Chung & Tam (2005)	Country: China (Hong Kong) Sample size: 30 (8 girls) Age: $\bar{x}$ = 10:4 years Setting: Special school Learning difficulties: IQ of 55-70 on WISC-III, around 70% on pre-test measure of word problem-solving, referred by teacher as needing assistance with word problems (all participants)	Research design: Intervention, Comparison 1 [taught to visualise problems] and Comparison 2 [extra regular teaching], all *n*=10 Participant assignment: Random Intervention content: All participants were given initial general instruction to ensure they understood the language used in the word problems, followed by 2 worked examples on a whiteboard and 3 problems on a worksheet, followed by feedback *Intervention* 5-step model based on CSI: (1) Read the problem aloud, (2) Select the important information, (3) Draw a representation of the problem, (4) Write down the steps, (5) Check the answer Three metacognitive strategies (say, ask, check) Participants were taught these strategies and received a reminder worksheet *Comparison 1* Participants were taught to visualise problems through the use of circles to demonstrate the concepts of addition and subtraction. Participants studied the worked solutions for three minutes before visualising their own problems by drawing circles Intervention delivery: Groups of 10 with the first author during regularly scheduled resource classes 4.16 hours (5 x 50 minute sessions)

Study	Participants	Procedure
Fuchs, Fuchs, Prentice, Burch, Hamlett, Owen, & Schroeter (2003)	Country: USA Sample size: 395 Age: 3rd grade Setting: Six mainstream schools Learning difficulties: Participants were designated as high- (HA, *n*=90), average- (AA, *n*=199), or low- (LA, *n*=106) achieving based on teacher consultation and the preceding year's district accountability testing	Research design: Intervention (*n*=137) [word problem-solving practice plus Self-regulated learning], Comparison 1 (*n*=138) [word problem-solving practice only], and Comparison 2 (*n*=120) [no extra teaching] Participant assignment: 24 teachers were randomly assigned across the three conditions (*n*=8 for each condition); all pupils in the class were in the same condition; each condition was represented approximately equally across the 6 schools; teacher groups were comparable on gender, age, education, and years teaching; student groups were comparable on gender, reduced or free lunch, race, SEN status, and English-as-a-second-language status Intervention content: Intervention and Comparison 1 participants had a 3-week introductory unit on basic maths problem-solving. They were taught rules for problem solution, given worked examples, partner and independent work, and homework. Sessions were organised into five units, with the introductory unit followed by four units on different types of problems. The final two sessions were a cumulative review. In addition, Intervention participants scored their independent work using an answer key that provided credit for the process and answer. They charted their daily scores, inspected them in each session and set goals to beat their high scores. They scored their homework using an answer key. They discussed how they had transferred the problem structures to other parts of the school day or outside of school. Intervention delivery: Research assistants modelled the first two sessions of each unit while teachers observed. Teachers delivered the remaining sessions. 18 hours (32 sessions; two per week for 16 weeks; first two sessions of each unit were 40 minutes, remaining sessions were 30 minutes)

Study	Participants	Procedure
Kajamies, Vaurus, & Kinnunen (2010)	Country: Finland Sample size: 429; there were 24 main participants (12 girls) and a large extra comparison group Age: 4th grade ($\bar{x}$ = 10:4 years) Setting: Twelve mainstream schools Learning difficulties: <32% on a pre-test measure of word problem-solving (24 main participants)	Research design: Intervention (*n*=8), Comparison 1 (*n*=8) [no extra teaching], Comparison 2 (*n*=8) [reading comprehension intervention], and Comparison 3 (*n*=405) [no extra teaching but represented the range of attainment levels in the schools] Participant assignment: For practical reasons, intervention participants were selected from 2 classes out of a total of 138 students across 12 schools who had low pre-test scores; pairwise-matched, same-sex controls were selected on the basis of problem-solving scores Intervention content: Researcher-designed computer game, Quest of the Silver Owl (Vaurus & Kinnunen, 2003), plus teacher discussion 6-step model: (1) read the problem carefully from beginning to end, (2) construct a representation of the problem, (3) decide how to solve the problem, (4) execute the necessary calculations, (5) interpret the outcome and formulate the answer, (6) evaluate the solution Whilst participants were solving problems in the computer game, the teacher scaffolded participants' engagement in and reflection on the model, questioned participants, modelled the steps, gave immediate and concrete feedback, explained the rationale of the model, and encouraged participants to make pictures of problems and activate prior knowledge Intervention delivery: Groups of two with the first author in a quiet room at school; half the sessions were delivered outside regular school hours 10.5 hours (14 x 45 minute sessions; 2 per week for 7 weeks) Comparison 2 participants had 20 hours of reading comprehension intervention from special teachers

Study	Participants	Procedure
Pennequin, Sorel, Nanty, & Fontaine (2010)	Country: France Sample size: 48 (25 girls) Age: 3rd grade ($\bar{x}$ = 8:10 years) Setting: Mainstream school Learning difficulties: Designated normal- or low-achieving (both *n*=24) in maths based on teacher report and previous test scores and confirmed by pre-test scores (no hard criteria)	Research design: Intervention and Comparison [extra regular teaching], both *n*=24 Participant assignment: Random Intervention content: Based on the Strategy Evaluation Matrix (Schraw, 1998). Strategies taught included skimming, slowing down, activating prior knowledge, mental integration, and diagrams. Instruction focussed on planning, monitoring and evaluating. Comparison participants received tuition on memory, reading, writing, and mathematical (arithmetic and geometry) activities Intervention delivery: Groups of 6 children with a Research Assistant 5 hours (5 x 60 minute sessions over the course of 7 weeks) The comparison group had the same delivery parameters

Study	Participants	Procedure
Teong (2003)	Country: Singapore Sample size: 40 Age: 11-12 years Setting: Mainstream school Learning difficulties: Designated low achieving in maths based on scores between 50-70% in a national test (all participants)	Research design: Intervention and Comparison [word problem-solving practice only] Participant assignment: Not stated Intervention content: All participants worked collaboratively in pairs on word problems using WordMath computer software (Looi & Tan, 1998) Intervention students were given direct explanation, explicit demonstration, and scaffolded practice in using a 5-step model, 'CRIME', (1) careful reading, (2) recall possible strategies, (3) implement possible strategies, (4) monitor, (5) evaluate They had a reminder card with the acronym Intervention delivery: Not stated who delivered the teaching 4 hours (4 x 60 minute sessions over the course of 2 weeks)

Study	Participants	Procedure
Wang, Fuchs, Fuchs, Gilbert, Krowka, & Abramson (2019)	Country: USA Sample size: 69 (36 girls) Age: 3rd grade Setting: Six Mainstream schools Learning difficulties: Either (a) performance below the 22nd percentile on the Wide Range Achievement Test–4 [WRAT-4] (Wilkinson & Robertson, 2006) or (b) WRAT-4 performance below the 31st percentile and a score less than three on the Second-Grade Calculations Battery–Minuends to 18 (Fuchs, Hamlett, et al., 2003) (all participants)	Research design: Intervention (*n*=23) [fractions and self-regulation teaching], Comparison 1 (*n*=24) [fractions teaching only], and Comparison 2 (*n*=26) [no extra teaching] Participant assignment: Random Intervention content: Intervention and Comparison 1 participants were taught using a teaching programme called Third-Grade Super Solvers. This included worked examples, thinking out loud, practice and feedback. Intervention participants received instruction and discussion of self-regulation topics such as self-sufficiency, partner support, goal setting, taking responsibility for planning learning, identifying strengths and weaknesses, and tracking progress. Teaching of self-regulation topics took 4-7 minutes per session. Intervention delivery: Small groups with trained tutors (research grant employees, not trained teachers) 22.75 hours (39 x 35 minute sessions; 3 per week for 13 weeks)

Study	Participants	Procedure
Zhu (2015)	Country: China Sample size: 150 (63 girls) Age: 4th grade Setting: Mainstream school Learning difficulties: Group 1: <25% on a municipality-level maths test but >25% on the reading test (*n*=16) Group 2: <25% on municipality-level maths and reading tests (*n*=19) Group 3: 45-70% on municipality-level maths and reading tests (*n*=74) Group 4: >70% on municipality-level maths and reading tests (*n*=41)	Research design: Intervention and Comparison [extra regular teaching], both *n*=75 Participant assignment: four classes in one school were randomly divided into two intervention and two comparison classes Intervention content: 7-step model (based on CSI): (1) reading the problem for understanding, (2) paraphrasing by putting the problem into one's own words, (3) visualising by drawing a schematic representation, (4) hypothesising or setting up a plan, (5) estimating or predicting the answer, (6) computing, and (7) checking that the plan and the answer are correct 3 metacognitive strategies (say, ask, check) All participants received word problem-solving classes additional to their regular classes In the comparison group teachers proceeded with 'business-as-usual' In the intervention group, word problems were modelled by the teachers, participants completed exercises with teacher scaffolding, the 7-step model and three strategies were displayed on the wall Intervention delivery: Trained teachers in four classrooms (Intervention teachers received two days' training from researchers, three mock lessons, and a set of CSI materials) Sessions delivered outside of regular classes 10.66 hours (16 x 40 minute sessions; two per week for eight weeks)

Participants

In total, 1161 participants took part in the reviewed studies, ranging from age 8 to twelve years. There was substantial variation in sample size, from 30 to 429. From available data, gender representation was roughly equal with 45% female participants (144/321). Fuchs et al. (2003) and Teong (2003) provided no data on gender, resulting in WoE C penalties. Wang et al. (2019) had attrition (7.25%) but this was low and balanced across conditions so was not penalised in WoE A.

Studies which used screening measures to identify low-achievers in maths were deemed more effective than those which used teacher report (due to issues of subjectivity and difficulty of replication) or national tests (because these did not necessarily provide up-to-date information) and were rated higher in WoE C. While it is difficult to compare participants' maths ability across studies because of study-specific means of testing, variation is likely. Some (Chung & Tam, 2005) required relatively high pre-test scores (70%), reasoning that children needed some maths knowledge to access the intervention. Others (Kajamies et al., 2010; Zhu, 2015) required low scores (<30%), reasoning that the lowest-achievers needed most help. Three studies (Fuchs et al., 2003; Pennequin et al., 2010; Zhu, 2015) had an additional independent variable of maths attainment, allowing for analysis of high-, normal-, and low-achievers.

Studies took place in China, Finland, France, Singapore, and USA. There was thus substantial heterogeneity in cultural background and educational systems among participants. This information was not appraised by the WoE framework because judgments could not be made by the author of the merits of different countries' education systems. However, it has positive implications for the generalisation of findings and potentially allows for cross-cultural analysis. A potential drawback for EPs working in the UK is the absence of UK-based evidence.

All studies sampled participants from mainstream schools except Chung and Tam (2005), who sampled from a special school. Studies which sampled from multiple schools (Fuchs et al.,

2003; Kajamies et al., 2010; Wang et al., 2019) received higher WoE C ratings because their findings generalised across institutions.

Studies which randomly assigned individual participants to groups received the highest WoE B ratings (Chung & Tam, 2005; Pennequin et al., 2010; Wang et al., 2019) because this minimises the chance of irrelevant, systematic between-groups differences (Barker et al., 2005). Small penalties were given to studies which randomly assigned classes of participants (Fuchs et al., 2003; Zhu, 2015) because of potential effects from previous shared learning experiences. Harsh penalties were given to Teong (2003), who did not state the sampling procedure, and Kajamies et al. (2010) for use of convenience sampling, which is open to selection bias and cannot be replicated. Furthermore, the latter study had unequal group sizes (three small groups (n=8) and an extra group (n=405)), which had ramifications for statistical analysis and led to penalties for WoE B, 'Power analysis'.

Research Design

All studies used an experimental design with pre-post testing and intervention/comparison groups. A variety of comparison groups were used. Fuchs et al. (2003), Teong (2003), and Wang et al. (2019) isolated the effect of metacognition instruction by including a comparison group which received equivalent teaching with equivalent delivery parameters minus the metacognitive components, leading to high WoE B ratings. Chung and Tam (2005), Pennequin et al. (2010), and Zhu (2015) had groups receiving additional regular maths teaching, while alternative interventions were provided by Chung and Tam (2005) (taught to visualise problems) and Kajamies et al. (2010) (reading comprehension instruction). These all controlled for attention effects (McCarney et al., 2007) but were less able to isolate the effect of metacognitive instruction, leading to lower WoE B ratings. Fuchs et al. (2003), Kajamies et al. (2010), and Wang et al. (2019) had groups receiving no additional teaching, representing the starkest contrast with participants receiving metacognitive instruction. Three studies included a second comparison group receiving regular teaching (Chung & Tam, 2005; Fuchs

et al., 2003; Wang et al., 2019), facilitating comparison of the metacognition intervention with both an alternative intervention and regular teaching.

Chung and Tam (2005) used pre-/post-measures of different levels of difficulty and did not analyse pre-/post-differences, resulting in a low score for WoE B, 'Outcome measures'. The pre-measure was used to establish between-groups similarity on problem-solving.

Studies which took follow-up measures (Chung & Tam, 2005; Kajamies et al., 2010; Teong, 2003) were rated higher in WoE C because this illustrated whether intervention benefits were maintained. Studies which measured other attributes of metacognition in addition to problem-solving scores (Fuchs et al., 2003; Kajamies et al., 2010; Pennequin et al., 2010; Teong, 2003) were rated higher in WoE C because this provided a more holistic picture and indicated whether participants could generalise learning. Studies which took 'far-transfer' measures (word-problems structured differently to those practised during intervention) were also rated higher in WoE C (Chung & Tam, 2005; Fuchs et al., 2003) because this indicated whether participants could apply learning in a novel mathematical context.

Intervention

Interventions ranged from four to 22.75 hours of total delivery time ($\bar{x}$ = 10.72, *SD* = 7.29) in four to 39 sessions lasting between 30 and 60 minutes between one and three times per week, indicating substantial heterogeneity. Interventions were delivered by researchers or research-assistants apart from Fuchs et al. (2003) and Zhu (2015), who trained teachers. This led to higher WoE C scores because it contributed to external validity, showing teachers with two days' training could deliver interventions. Teong (2003) did not state who delivered the intervention, hindering replicability.

Content and procedures of metacognition instruction differed. The protocol operationalising metacognition (see Table B9) revealed different focus points in terms of person, task, and strategy knowledge (Flavell, 1979). No study taught all areas of metacognitive knowledge. All

studies combined teacher instruction with independent practice. Kajamies et al. (2010) and Teong (2003) had participants practising on computers while others practised on paper.

Four studies provided participants with sequential problem-solving models; three (Chung & Tam, 2005; Kajamies et al., 2010; Zhu, 2015) were derived from CSI and one (Teong, 2003) was researcher-developed but similar. These studies focused on strategy knowledge, particularly the plan-monitor-evaluate sequence. Only Kajamies et al. (2010) addressed person knowledge – engaging participants in peer discussion – and task knowledge – deciding which strategy was appropriate for each task (also addressed by Teong, 2003). Pennequin et al. (2010) adopted a similar strategic focus without provision of a problem-solving model. Person and task knowledge were addressed as by Kajamies et al. (2010) but participants were not given mathematical problem-solving teaching.

Fuchs et al. (2003) and Wang et al. (2019) adopted person-focused teaching, labelling their interventions 'self-regulated learning'. There was a focus on analysing participants' strengths and weaknesses through goal-setting, marking and evaluating work, and tracking progress. Fuchs et al. (2003) were the only researchers to discuss with participants how they had transferred learning to other subjects or areas outside of school. Since metacognitive skills are potentially domain-general (Schraw, 1998), applying skills beyond the intervention context is likely a helpful learning process. Whereas Fuchs et al. (2003) embedded their procedures throughout the teaching sessions, Wang et al. (2019) had 4-7-minute self-regulation 'sections' in each session.

Fuchs et al. (2003), Wang et al. (2019), and Zhu (2015) assessed fidelity of intervention implementation, contributing to WoE A. This is important for replicability and generalisation because it shows whether the intervention described in the article was carried out accurately. Fuchs et al. (2003) reviewed audiotapes against a checklist of key information points, averaging 96% (*SD* = 6.91) in the intervention and 96.9% (*SD* = 9.56) in the comparison group. Wang et al. (2019) also reviewed audiotapes, averaging 95% (*SD* = 12.86) of self-regulation learning points covered. Zhu (2015) observed teachers in lessons, averaging 90% of lesson

components in the intervention and 93% in the comparison group. While it is laudable for these studies to have assessed fidelity, their methods for doing so were limited to procedural checklists of teaching points. Other factors to consider include the instructors' understanding of the intervention, behavioural aspects of the instructors' delivery (e.g. clarity of expression), and student engagement with instructors (Stains & Vickrey, 2017).

Findings

Table 8 presents a summary of primary and secondary outcome measures, descriptions of findings, and effect sizes. The primary measure was score on a word problem-solving test. Secondary measures varied across studies.

The effect size calculated for all studies was the standardised mean difference (Cohen's *d*). Where possible, this was calculated by the author as the difference between intervention and comparison improvement (post-test minus pre-test) means divided by the pooled standard deviation of pre-test means (Morris, 2008). If there were insufficient data, only post-test means were used. Pennequin et al. (2010) provided no descriptive statistics so effect sizes were calculated using the *F*-statistic of the interaction between pre-/post-scores and intervention/comparison with the Campbell Collaboration online calculator (Wilson, n.d.). Effect size descriptors are provided in Table 7.

A meta-analysis was conducted to assess the overall effect of metacognition instruction on problem-solving. The comparison groups in the meta-analysis were those which provided the biggest experimental contrast from each study (a group receiving no additional teaching or extra regular teaching). While this does not allow consideration of the most effective way to deliver metacognition instruction, it gives an average baseline figure against which future meta-analyses could compare (Law et al., 2004). The meta-analysis was conducted with a random-effects model using *Meta-Essentials* software (Suurmond et al., 2017). A random-effects model was preferred over a fixed-effects model because of the substantial heterogeneity between studies in terms of participants and intervention, meaning there are

likely to be a range of 'true' effect-sizes (Borenstein et al., 2010). The software converted Cohen's *d* to Hedge's *g*. A summary of effect sizes and meta-analysis parameters is provided in Table 9. A Forest plot is provided in Figure 4.

The combined effect size across seven studies was *g*=1.39 (95% confidence intervals (CIs) [0.73, 2.04]). This can be described as large, with a medium-large lower confidence interval. This statistical evidence is supported by WoE D ratings. Fuchs et al. (2003) were rated High and had the largest effect size, while Wang et al. (2019) and Zhu (2015) were rated Medium-High and had large effect sizes. These studies were all adequately powered. Only Teong (2003) got a Low rating. The average WoE D rating across studies was 1.95 (*SD* = .5). It was hypothesised that effect size may be related to length of intervention, but this correlation was weak-moderate, *r*=.361.

Visual analysis of a funnel plot (see Figure 5) shows a cluster of small-*n* studies with lower effect sizes and one very large effect size (Fuchs et al., 2003), which lays outside the funnel. With so few studies, it is hard to draw conclusions from the forest plot. However, when the meta-analysis was re-run omitting the apparent outlier, the combined effect size (1.19, 95% CIs [0.75, 1.62]) did not substantially change so there were no considerations of practical significance.

A second meta-analysis was conducted using comparison groups that received the same instruction as intervention groups minus the metacognitive components. This attempted to isolate the effect of metacognition instruction from all other effects of intervention including increased attention from teachers, problem-solving practice, and the novel experience of taking part in research. A summary is provided in Table 10. A more conservative picture emerged with a combined effect size of *g*=0.36 (95% CIs [-0.65, 1.37]), which is small. Since the lower CI crosses 0, there is only poor statistical evidence that metacognition instruction provided benefit above and beyond the other components of intervention.

Table 7

Descriptors for Cohen's d *and Hedge's* g (Cohen, 1992)

Effect size	Descriptor
.8	Large
.5	Medium
.2	Small

Table 8

Outcomes, Main Findings, and Selected Effect Sizes (ES; Cohen's d*; 2 Decimal Places) from the Seven Included Studies*

Study	Outcomes	Main findings	ES Description	ES	WoE D
Chung & Tam (2005)	Content: Eight 2-step addition and subtraction word problems modified according to participants' prior knowledge from textbooks; five problems were similar to previously-encountered examples, three were presented in a novel format Collection: pre-test, post-test, follow-up 14 days post-intervention	Pre-test scores were not included in statistical analysis so no results were reported for pre- to post-test differences for any groups (the three groups received similar pre-test scores, ranging from 66.2% to 67.8%) Intervention (I) and Comparison 1 (C1) scored significantly higher at post-test and follow-up than Comparison 2 (C2) but did not differ from each other Post-test scores were maintained at follow-up by both I and C1 but not by C2	Calculation: Post-test, comparison vs intervention, M & SD I vs C1 (active, taught to visualise problems) I vs C2 (active, extra regular teaching) C1 vs C2	 -0.09 (small) 1.17 (large) 1.40 (large)	1.6 Medium

Study	Outcomes	Main findings	ES Description	ES	WoE D
Fuchs, Fuchs, Prentice, Burch, Hamlett, Owen, & Schroeter (2003)	Content: 10 *immediate transfer* word problems (four problem structures with novel cover stories), seven *near transfer* problems (four problem structures, novel cover stories, one superficial problem feature varied), one *far transfer* problem (all structures embedded in a real-life context, with all superficial features varied and elements of novelty) Collection: pre-test, post-test Secondary outcomes: participant questionnaire of self-regulation processes	Pre-test scores did not differ between treatment groups other than the differences between ability groupings Immediate transfer - C2 improvement was less than C1, which in turn was less than I; this was found across low, average and high achieving participants Near transfer – average and low achieving participants in C1 and I improved more than C 2 but did not differ from each other Far transfer – C2 improvement was less than C1, which in turn was less than I; this was found across low, average and high achieving participants	Calculation: Pre- vs post-test, comparison vs intervention, M & SD (low-achievers only) *Immediate transfer* I vs C1 (active, word problem-solving practice only) I vs C2 (passive) C1 vs C2 *Near transfer* I vs C1 I vs C2 C1 vs C2 *Far transfer* I vs C1 I vs C2 C1 vs C2	 0.33 (small) 2.68 (large) 1.83 (large) 0.35 (small) 2.18 (large) 1.24 (large) 0.21 (small) 1.17 (large) 0.69 (medium)	2.5 High

Study	Outcomes	Main findings	ES Description	ES	WoE D
Kajamies, Vaurus, & Kinnunen (2010)	Content: Fifteen one-step and multi-step word problems; maximum score 86 Collection: pre-test, post-test, follow-up 6 months post-intervention Secondary outcomes: teacher ratings of task orientation, researcher-designed measure of arithmetical skills, standardised test of non-verbal intelligence (Raven et al., 2000), national test of reading comprehension	All four groups increased their scores at both post-test and follow-up I participants' scores increased from pre- to post-test significantly more than C3 participants' scores At follow-up, Intervention participants' scores no longer significantly differed from C3 participants' scores (but C1 and C2 were still lower than both) C1 and C2 participants' scores did not increase at a differential rate Individual analysis of I participants showed that 3/8 had large improvements from pre- to post-test, 2/8 had no improvement, and 3/8 had lower post-test scores	Calculation: Pre- vs post-test, comparison vs intervention, M & SD I vs C1 (passive) I vs C2 (active, reading comprehension)	 0.74 (medium-large) 0.67 (medium)	1.72 Medium

Study	Outcomes	Main findings	ES Description	ES	WoE D
Pennequin, Sorel, Nanty, & Fontaine (2010)	Content: 12 word problems of varying difficulty Collection: pre-test, post-test Secondary outcomes: participant ratings of the importance of metacognitive knowledge propositions; participant predictions of problem-solving test performance	Pre-test scores did not differ between groups Both normal- and low-achievers in the Intervention had higher post-test scores but neither Comparison group had higher post-test scores I low-achievers improved more than normal-achievers At post-test there was no longer a difference in scores between I low- and normal-achievers On a measure of metacognitive knowledge, only low-achievers had increased post-test scores	Calculation: Pre- vs post-test, comparison vs intervention, *F* value of Pre/post x Group interaction [a] I vs C (active, extra regular teaching)	1.21 (large)	1.9 Medium

Study	Outcomes	Main findings	ES Description	ES	WoE D
Teong (2003)	Content: Ten multi-step word problems Collection: pre-test, post-test, follow-up six weeks post-intervention Secondary outcomes: observations of cognitive and metacognitive behaviours from video recordings of participants thinking aloud	Pre-test scores differed between groups so post-test and follow-up scores were adjusted with the pre-test scores used as a covariate Intervention scores increased from pre- to post-test and increased further at follow-up Post-test Intervention scores were significantly higher than Comparison scores	Calculation: Pre- vs post-test, comparison vs intervention, M & SD I vs C (post-adjustment) (active, word problem-solving practice only)	0.91 (large)	1.15 Low

Study	Outcomes	Main findings	ES Description	ES	WoE D
Wang, Fuchs, Fuchs, Gilbert, Krowka, & Abramson (2019)	Content: 16 fraction word problems from the Fraction Battery-Revised-Fraction Word Problems (Schumacher et al., 2015); maximum score 36 Collection: pre-test, post-test Secondary outcomes: other sub-tests from the Fraction Battery – Revised: single-digit multiplication, comparing fractions, ordering fractions, and number line 0-1; broad fraction understanding from a national measure	Intervention post-test scores were higher than C2 C1 post-test scores were also higher than C2 but there was a moderation effect of pre-test scores; C1 participants with higher pre-test scores responded more adequately to the fractions teaching than those with lower pre-test scores This moderation effect was not apparent for Intervention	Calculation: Post-test, comparison vs intervention, M & SD I vs C1 (active, fractions teaching only) I vs C2 (passive) C1 vs C2	 -0.04 (small) 1.00 (large) 0.91 (large)	2.43 Medium-High

Study	Outcomes	Main findings	ES Description	ES	WoE D
Zhu (2015)	Content: 12 word problems of four different types selected from textbooks; scores from 0-3 possible; scored by research assistants Collection: one week pre-test, one week post-test	Pre-test, there were no significant differences between respective intervention and comparison groups	Calculation: Pre- vs post-test, comparison vs intervention, M & SD		2.37 Medium-High
		All four intervention groups outperformed their respective comparison groups	I vs C (Maths Difficulties only) (active, extra regular teaching)	1.82 (large)	
		Intervention groups all had significant improvement from pre- to post-test	I vs C (MD & Reading Difficulties)	0.99 (large)	
		Comparison low-achievers showed no or little response to regular teaching	I vs C (Average Achieving)	1.48 (large)	
		Intervention low-achievers did not have significantly different pre-test scores but group 1 significantly out-performed group 2 on the post-test. Participants with only low maths scores benefitted more from the intervention than those with low maths and reading scores	I vs C (High Achieving)	0.72 (medium -large)	
		Group 4 (high-achievers) benefitted less from the intervention than the other three groups			

[a] Due to Pennequin et al. (2010) not reporting descriptive statistics, the *F* value of the interaction between pre-/post-score and experimental group provided the best effect size estimate. However, this meant conflating the 'normal' and 'low' achievers' scores. Thus, the reported effect size is an underestimate of the improvement of the low-achievers in the intervention group.

Table 9

Sample Sizes of Intervention (n_1) and Comparison (n_2) Groups, Converted Effect Sizes (ES; Hedges' g (Standard Error (SE))), 95% Confidence Intervals (95% CIs), Weight, and the Combined ES of the Seven Studies Included in the First Meta-Analysis

Study	n_1	n_2	ES (SE)	95% CIs	Weight
Chung & Tam (2005)	10	10	1.12 (0.46)	0.20, 2.14	12.39%
Fuchs et al. (2003)	137	120	2.67 (0.17)	2.34, 3.02	16.10%
Kajamies et al. (2010)	8	8	0.70 (0.49)	-0.31, 1.78	12.04%
Pennequin et al. (2010)	24	24	1.19 (0.31)	0.59, 1.83	14.55%
Teong (2003)	20	20	0.89 (0.33)	0.25, 1.56	14.33%
Wang et al. (2019)	23	26	0.98 (0.30)	0.40, 1.60	14.68%
Zhu (2015)	75	75	1.81 (0.19)	1.44, 2.20	15.90%
Combined	580		1.39 (0.27)	0.73, 2.04	

Figure 4

Forest Plot of Effect Sizes (95% Confidence Intervals) of the Seven Included Studies and the Combined Effect Size from the First Meta-Analysis

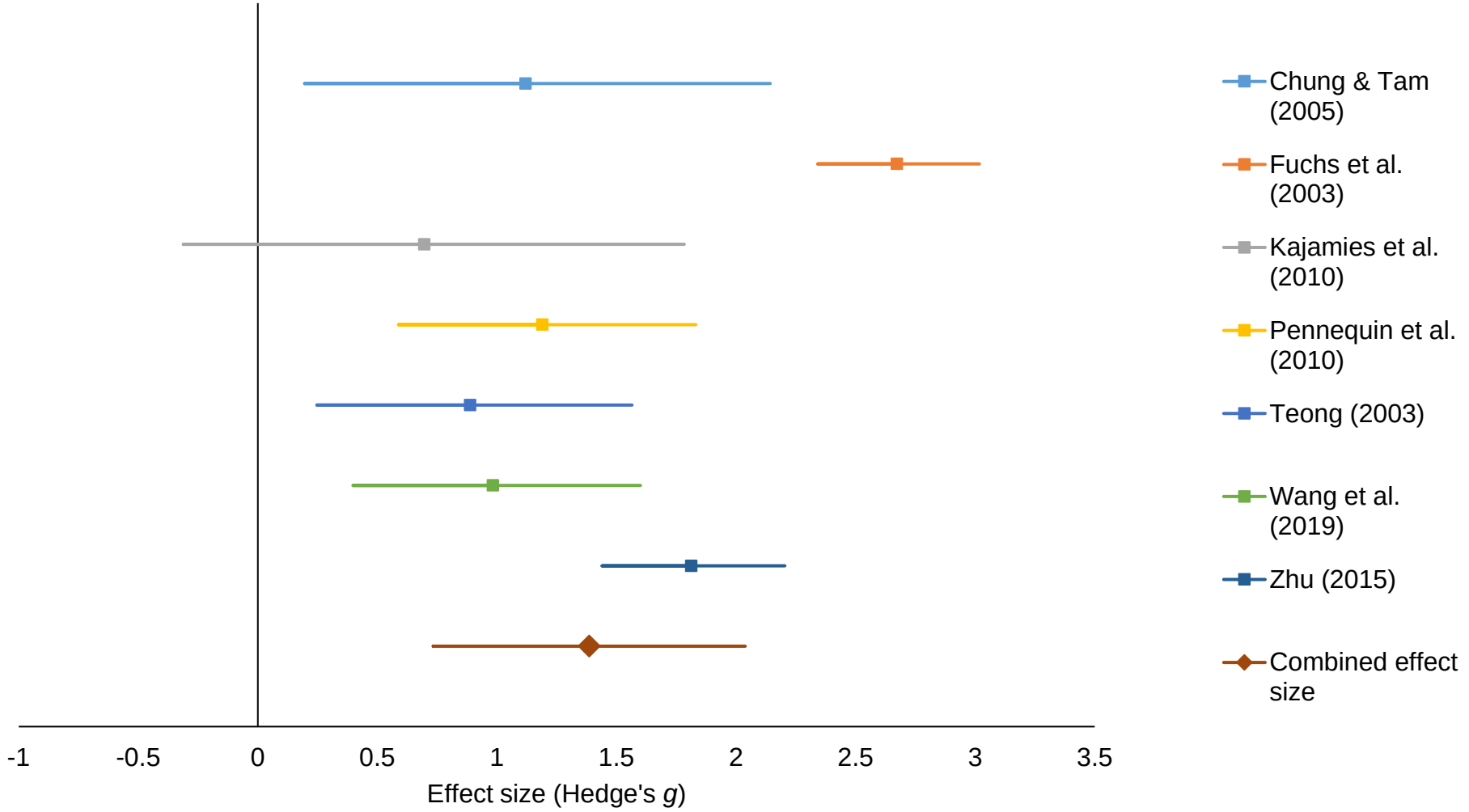


Figure 5

Funnel Plot of Effect Sizes from Individual Studies and the Combined Effect Size (95% Confidence Intervals) Against Standard Error from the First Meta-Analysis

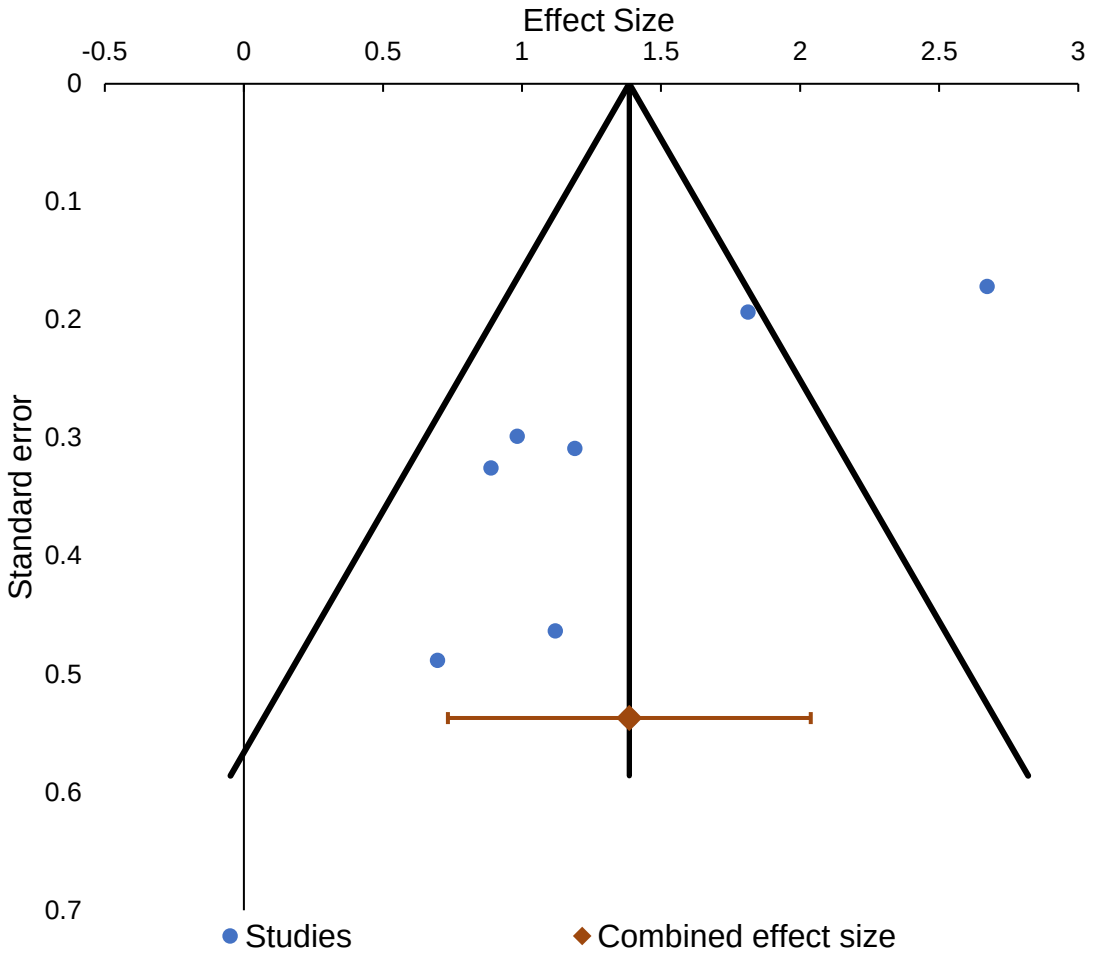


Table 10

Sample Sizes of Intervention (n_1) and Comparison (n_2) Groups, ES (Hedges' g (Standard Error (SE))), 95% Confidence Intervals (95% CIs), Weight, and the Combined ES of the Three Studies Included in the Second Meta-Analysis

Study	n_1	n_2	ES (SE)	95% CIs	Weight
Fuchs et al. (2003)	137	138	0.33 (0.12)	0.09, 0.57	48.74%
Teong (2003)	20	20	0.89 (0.33)	0.25, 1.56	23.82%
Wang et al. (2019)	23	24	-0.04 (0.29)	-0.61, 0.54	27.44%
Combined	362		0.36 (0.23)	-0.65, 1.37	

Conclusions and Recommendations

Discussion of Findings

This review evaluated whether metacognition instruction improved word problem-solving of children who were low-achievers in maths. Seven studies met the inclusion criteria, with one receiving a High WoE D rating, two Medium-High, three Medium, and one Low.

Given the combined evidence of statistical effect, methodological quality, and methodological and topical relevance, it can be concluded that interventions incorporating metacognition instruction had a considerable effect on problem-solving. Gersten et al. (2005) suggested, for an intervention to be 'evidence-based practice', there should be two studies with High WoE A (three were found in this review) and a combined effect size significantly greater than zero. This review supports the claim for interventions incorporating metacognition instruction as evidence-based practice.

Evidence for the unique contribution of metacognition instruction above other intervention components is equivocal. Based on the second meta-analysis, it cannot be confidently concluded that there was an effect on problem-solving. However, studies with follow-up measures (Chung & Tam, 2005; Kajamies et al., 2010; Teong, 2003) found intervention participants maintained gains to a greater degree than comparison participants. Metacognition instruction may promote longer-term learning but it is difficult to assess given the lack of studies with follow-up measures and strong methodologies. Information from secondary outcomes is potentially enlightening. Fuchs et al. (2003) found through a questionnaire that intervention participants self-rated as having higher self-efficacy (d = 0.92), goal orientation and self-monitoring (d = 1.2) than comparison participants who had the same intervention minus metacognitive components. Wang et al. (2019) found through a distal measure of general fraction tasks that intervention participants scored higher (d = 0.44) than comparison participants who had the same intervention minus metacognitive components. These findings tentatively support the theoretical claim that metacognitive knowledge and skills generalise

beyond domains (Schraw, 1998), which may be a unique, additional contribution to conventional interventions.

Recommendations for Practice

When considering the appropriateness of an intervention for educational psychology practice, evaluation of generalisability is key. The two studies which trained teachers to implement interventions (Fuchs et al., 2003; Zhu, 2015) had High or Medium-High WoE D ratings and large effect sizes, suggesting teacher delivery is feasible. Furthermore, three studies (including the aforementioned) sampled from multiple mainstream or special schools, suggesting results generalised across settings. In the absence of a commercial intervention package, the only cost of metacognition instruction is teacher training, either in CSI or general metacognitive principles. This is likely to have significant returns because teachers could utilise knowledge in classrooms and interventions.

In terms of participant characteristics, generalisability is less clear. Results did replicate across culturally disparate populations with different school structures. However, no studies took place in the UK. It may be inferred from evidence of cross-cultural replication that similar results would be found with a UK sample but this cannot be assumed.

Overall, however, given the substantial benefits of metacognition instruction for children who are low-achievers in maths, and the simplicity and low cost of its implementation, it should be recommended by educational psychologists.

Limitations of the Review

It could be argued this review's inclusion criteria permitted studies which taught cognitive as well as metacognitive strategies. However, given the domain-specific origins of metacognition (Schraw, 1998) it would seem conceptually and developmentally inconsistent to teach metacognition in isolation without any relevance to a particular subject, particularly for young learners who are struggling. Therefore, a review of studies which only taught metacognition would have had weaker external validity for educational psychology practice even if it provided

stronger theoretical evidence. Furthermore, an attempt was made to isolate the effect of metacognition through a secondary meta-analysis.

This review's development of a protocol to operationalise metacognition for WoE C may be questioned. Its current form privileges comprehensiveness over quality (although quality is assessed elsewhere) and has not been tested for construct or inter-rater reliability. However, given the finding that metacognitive knowledge and regulation contributed separately to test performance (Schraw & Dennison, 1994), there is evidence that teaching multiple metacognitive components leads to better outcomes. Furthermore, protocol development was necessary given the lack of an existing protocol in the field.

Recommendations for Future Research

Previous authors have noted the need to identify which components of metacognition are important in facilitating change (Dowker, 2017). While this review did not address this question it did illustrate a dual focus in the literature on either strategy knowledge or self-regulation. Future research could explore the differential effect of these focus points and whether effects are additive.

Future studies should include follow-up measures and secondary outcomes to evaluate maintenance and generalisability. This is important given the domain-general nature of metacognition (Schraw, 1998) and its potential cross-curricular impact.

References

Barker, C., Pistrang, N., & Elliott, R. R. (2005). *Research methods in clinical psychology: An introduction for students and practitioners* (2nd ed.). Wiley-Blackwell.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Chung, K. K. H., & Tam, Y. H. (2005). Effects of cognitive-based instruction on mathematical problem solving by learners with mild intellectual disabilities. *Journal of Intellectual and Developmental Disability*, *30*(4), 207–216. https://doi.org/10.1080/13668250500349409

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

De Corte, E., Verschaffel, L., & Eynde, P. O. (2000). Self-regulation: A characteristic and a goal of mathematics education. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 687–726). Academic Press. https://doi.org/http://dx.doi.org/10.1016/B978-012109890-2/50050-0

Dowker, A. (2017). Interventions for primary school children with difficulties in mathematics. In J. Sarama, D. H. Clements, C. Germeroth, & C. DayHess (Eds.), *Development of early childhood mathematics education* (Vol. 53, pp. 255–287). https://doi.org/10.1016/bs.acdb.2017.04.004

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/BF03193146

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1002/bit.23191

Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., & Schroeter, K. (2003). Enhancing third-grade students' mathematical problem solving with self-regulated learning strategies. *Journal of Educational Psychology*, *95*(2), 306–315. https://doi.org/10.1037/0022-0663.95.2.306

Fuchs, L. S., Hamlett, C. L., & Powell, S. R. (2003). *Second-grade calculations battery*.

Garner, R. (1987). *Metacognition and reading comprehension*. Ablex Publishing.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*(2), 149–164. https://doi.org/10.1177/001440290507100202

Gleitman, L. R., Gleitman, H., & Shipley, E. F. (1972). The emergence of the child as grammarian. *Cognition*, *1*(2–3), 137–164. https://doi.org/10.1016/0010-0277(72)90016-9

Gough, D. (2007). Weight of evidence: A framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education*, *22*(2), 213–228. https://doi.org/10.1080/02671520701296189

Kajamies, A., Vauras, M., & Kinnunen, R. (2010). Instructing low-achievers in mathematical word problem solving. *Scandinavian Journal Of Educational Research*, *54*(4), 335–355. https://doi.org/10.1080/00313831.2010.493341

Krathwohl, D. R. (2002). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. *Theory into Practice*, *41*(4), 212–218.

Law, J., Garrett, Z., & Nye, C. (2004). The efficacy of treatment for children with developmental speech and language delay/disorder: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *47*(4), 924–943. https://doi.org/10.1044/1092-4388(2004/069)

Livingston, J. A. (1996). *Effects of metacognitive instruction on strategy use of college students [Unpublished manuscript]*. State University of New York at Buffalo.

Looi, C. K., & Tan, B. T. (1998). A cognitive-apprenticeship-based environment for learning word problem solving. *Journal of Computers in Mathematics and Science Teaching*, *17*(9), 339–354.

McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M., & Fisher, P. (2007). The Hawthorne Effect: A randomised, controlled trial. *BMC Medical Research Methodology*, *7*. https://doi.org/10.1186/1471-2288-7-30

Miller, S. P., & Mercer, C. D. (1997). Educational aspects of mathematics disabilities. *Journal of Learning Disabilities*, *30*(1), 47–56.

Montague, M. (1997). Cognitive strategy instruction in mathematics for students with learning disabilities. *Journal of Learning Disabilities*, *30*(2), 164–177. https://doi.org/10.1177/002221949703000204

Montague, M., & Dietz, S. (2009). Evaluating the evidence base for cognitive strategy instruction and mathematical problem solving. *Exceptional Children*, *75*(3), 285–302. https://doi.org/10.1177/001440290907500302

Montague, M., Enders, C., & Dietz, S. (2011). Effects of Cognitive Strategy Instruction on Math Problem Solving of Middle School Students With Learning Disabilities. *Learning Disability Quarterly*, *34*(4), 262–272. https://doi.org/10.1177/0731948711421762

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*(2), 364–386. https://doi.org/10.1177/1094428106291059

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

Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: Horses for courses. *Journal of Epidemiology and Community Health*, *57*, 527–529.

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

Quigley, A., Muijs, D., & Stringer, E. (2018). Metacognition and Self-regulated Learning: Guidance Report. In *The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit*.

Raven, J., Raven, J. C., & Court, J. H. (2000). *Standard progressive matrices*. Psychologist Press.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, *26*(1–2), 113–125. https://doi.org/10.1007/978-94-017-2243-8_1

Schraw, G., & Dennison, R. S. (1994). Assessing Metacognitive Awareness. *Contemporary Educational Psychology*, *19*(4), 460–475. https://doi.org/10.1006/ceps.1994.1033

Schumacher, R. F., Namkung, J., Malone, A., Wang, A., Abramson, R., & Fuchs, L. (2015). *Fraction Battery - Revised*.

Stains, M., & Vickrey, T. (2017). Fidelity of implementation: An overlooked yet critical construct to establish effectiveness of evidence-based instructional practices. *CBE: Life Sciences Education*, *16*(1), 1–11. https://doi.org/10.1187/cbe.16-03-0113

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

Teong, S. K. (2003). The effect of metacognitive training on mathematical word-problem solving. *Journal of Computer Assisted Learning*, *19*(1), 46–55. https://doi.org/10.1046/j.0266-4909.2003.00005.x

Vaurus, M., & Kinnunen, R. (2003). *Quest of the silver owl, teaching game for improving arithmetic and mathematical problem solving skills*. University of Turku, Centre for Learning Research.

Veenman, M. V. J. (2015). Metacognition. In P. Afflerbach (Ed.), *Handbook of Individual Differences in Reading: Reader, Text, and Context*. Routledge. https://doi.org/10.4324/9780203075562.ch3

Veenman, M. V. J., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences*, *15*(2), 159–176. https://doi.org/10.1016/j.lindif.2004.12.001

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, M. C., Haertel, G. D., & Walberg, H. J. (1990). What Influences Learning? A Content Analysis of Review Literature. *Journal of Educational Research*, *84*(1), 30–43. https://doi.org/10.1080/00220671.1990.10885988

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide range achievement test* (4th ed.). Wide Range.

Wilson, D. B. (n.d.). *Campbell Collaboration Effect Size Calculator*. Retrieved January 17, 2020, from http://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD3.php

Zhu, N. (2015). Cognitive strategy instruction for mathematical word problem-solving of students with mathematics disabilities in China. *International Journal of Disability, Development and Education*, *62*(6), 608–627. https://doi.org/10.1080/1034912X.2015.1077935

Appendix A

Excluded articles

A list of references of studies excluded at full text screening is provided in Table A1. References are not provided for studies excluded at title and abstract screening but the total number of studies excluded under each criterion is shown in Figure 3.

Table A1

Articles Excluded at Full Text Screening with Reference to Exclusion Criteria

Excluded study	Exclusion criterion
Bishara, S. (2016). Self-regulated math instructions for pupils with learning disabilities. *Cogent Education*, *3*(1).	12
Björn, P. M., Äikäs, A., Hakkarainen, A., Kyttälä, M., & Fuchs, L. S. (2019). Accelerating mathematics word problem-solving performance and efficacy with think-aloud strategies. *South African Journal of Childhood Education*, *9*(1), 1-10.	5
Carcoba Falomir, G. A. (2019). Diagramming and algebraic word problem solving for secondary students with learning disabilities. *Intervention in School and Clinic*, *54*(4), 212-218.	6
Cardelle-Elawar, M. (1992). Effects of teaching metacognitive skills to students with low mathematics ability. *Teaching and teacher education*, *8*(2), 109-121.	12
Cardelle-Elawar, M. (1995). Effects of metacognitive instruction on low achievers in mathematics problems. *Teaching and Teacher Education*, *11*(1), 81-95.	12
Case, L. P., Harris, K. R., & Graham, S. (1992). Improving the mathematical problem-solving skills of students with learning disabilities: Self-regulated strategy development. *The Journal of Special Education*, *26*(1), 1-19.	6
Cassel, J., & Reid, R. (1996). Use of a self-regulated strategy intervention to improve word problem-solving skills of students with mild disabilities. *Journal of Behavioral Education*, *6*(2), 153-172.	6
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.	8

Excluded study	Exclusion criterion
Coughlin, J., & Montague, M. (2011). The effects of cognitive strategy instruction on the mathematical problem solving of adolescents with spina bifida. *The Journal of Special Education*, *45*(3), 171-183.	6
Crawford, L., Higgins, K. N., & Freeman, B. (2012). Exploring the use of active electronic support tools by students with learning disabilities. *Learning Disabilities: A Multidisciplinary Journal*, *18*(3), 135-144.	6
Cuenca-Carlino, Y., Freeman-Green, S., Stephenson, G. W., & Hauth, C. (2016). Self-regulated strategy development instruction for teaching multi-step equations to middle school students struggling in math. *The Journal of Special Education*, *50*(2), 75-85.	6
Desoete, A., Roeyers, H., & De Clercq, A. (2003). Can offline metacognition enhance mathematical problem solving?. *Journal of Educational Psychology*, *95*(1), 188.	11
Ennis, R. P., & Losinski, M. (2019). SRSD Fractions: Helping students at risk for disabilities add/subtract fractions with unlike denominators. *Journal of Learning Disabilities*, *52*(5), 399-412.	6
Fuchs, L. S., & Fuchs, D. (2005). Enhancing mathematical problem solving for students with disabilities. *The Journal of Special Education*, *39*(1), 45-57.	4
Hacker, D. J., Kiuhara, S. A., & Levin, J. R. (2019). A metacognitive intervention for teaching fractions to students with or at-risk for learning disabilities in mathematics. *ZDM*, *51*(4), 601-612.	4
Hua, Y., Morgan, B. S., Kaldenberg, E. R., & Goo, M. (2012). Cognitive strategy instruction for functional mathematical skill: Effects for young adults with intellectual disability. *Education and Training in Autism and Developmental Disabilities*, *47*(3), 345-358.	7
Hutchinson, N. L. (1993). Effects of cognitive strategy instruction on algebra problem solving of adolescents with learning disabilities. *Learning Disability Quarterly*, *16*(1), 34-63.	6
Iseman, J. S., & Naglieri, J. A. (2011). A cognitive strategy instruction to improve math calculation for children with ADHD and LD: A randomized controlled study. *Journal of Learning Disabilities*, *44*(2), 184-195.	11
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.	11
Joseph, L. M., & Hunter, A. D. (2001). Differential application of a cue card strategy for solving fraction problems: Exploring instructional utility of	6

Excluded study	Exclusion criterion
the cognitive assessment system. *Child Study Journal*, *31*(2), 123-137.	
Karbasdehi, E. R., Abolghasemi, A., & Khanzadeh, A. A. H. (2019). The effect of self-regulation empowerment program training on neurocognitive and social skills in students with dyscalculia. *Archives of Psychiatry and Psychotherapy*, *2*, 71-80.	7
Kiuhara, S. A., Gillespie Rouse, A., Dai, T., Witzel, B. S., Morphy, P., & Unker, B. (2019). Constructing written arguments to develop fraction knowledge. *Journal of Educational Psychology*. Advanced online publication.	7
Krawec, J., Huang, J., Montague, M., Kressler, B., & Melia de Alba, A. (2013). The effects of cognitive strategy instruction on knowledge of math problem-solving processes of middle school students with learning disabilities. *Learning Disability Quarterly*, *36*(2), 80-92.	7
Lazakidou, G., & Retalis, S. (2010). Using computer supported collaborative learning strategies for helping students acquire self-regulated problem-solving skills in mathematics. *Computers & Education*, *54*(1), 3-13.	11
Lucangeli, D., Fastame, M. C., Pedron, M., Porru, A., Duca, V., Hitchcott, P. K., & Penna, M. P. (2019). Metacognition and errors: The impact of self-regulatory trainings in children with specific learning disabilities. *ZDM*, *51*(4), 577-585.	7
Maras, K., Gamble, T., & Brosnan, M. (2019). Supporting metacognitive monitoring in mathematics learning for young people with autism spectrum disorder: A classroom-based study. *Autism*, *23*(1), 60-70.	5
Montague, M. (1992). The effects of cognitive and metacognitive strategy instruction on the mathematical problem solving of middle school students with learning disabilities. *Journal of Learning Disabilities, 25*(4), 230-248.	6
Montague, M., & Bos, C. S. (1986). The effect of cognitive strategy training on verbal math problem solving performance of learning disabled adolescents. *Journal of Learning Disabilities*, *19*(1), 26-33.	6
Montague, M., Applegate, B., & Marquard, K. (1993). Cognitive strategy instruction and mathematical problem-solving performance of students with learning disabilities. *Learning Disabilities Research & Practice, 8*(4), 223-232.	6
Naglieri, J. A., & Johnson, D. (2000). Effectiveness of a cognitive strategy intervention in improving arithmetic computation based on the PASS theory. *Journal of Learning Disabilities*, *33*(6), 591-597.	6

Excluded study	Exclusion criterion
Otto, B., & Kistner, S. (2017). Is there a Matthew effect in self-regulated learning and mathematical strategy application? Assessing the effects of a training program with standardized learning diaries. *Learning and Individual Differences*, *55*, 75-86.	6
Özsoy, G. (2011). An investigation of the relationship between metacognition and mathematics achievement. *Asia Pacific Education Review*, *12*(2), 227-235.	5
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), 17-31.	11
Pfannenstiel, K. H., Bryant, D. P., Bryant, B. R., & Porterfield, J. A. (2015). Cognitive strategy instruction for teaching word problems to primary-level struggling students. *Intervention in school and clinic*, *50*(5), 291-296.	6
Salihu, L., Aro, M., & Räsänen, P. (2017). Dynamic potential of feedback in self-regulated learning and motivation of children with mathematical learning difficulties. *Hrvatska revija za rehabilitacijska istraživanja*, *53*(2), 111-129.	5
Seo, Y. J., & Bryant, D. (2012). Multimedia CAI program for students with mathematics difficulties. *Remedial and Special Education*, *33*(4), 217-225.	6
Sheriff, K. A., & Boon, R. T. (2014). Effects of computer-based graphic organizers to solve one-step word problems for middle school students with mild intellectual disability: A preliminary study. *Research in Developmental Disabilities*, *35*(8), 1828-1837.	5
Shin, M., & Bryant, D. P. (2017). Improving the fraction word problem solving of students with mathematics learning disabilities: Interactive computer application. *Remedial and Special Education*, *38*(2), 76-86.	6
Tzohar-Rozen, M., & Kramarski, B. (2014). Metacognition, motivation and emotions: Contribution of self-regulated learning to solving mathematical problems. *Global Education Review*, *1*(4).	11
Tzohar-Rozen, M., & Kramarski, B. (2017). Meta-cognition and meta-affect in young students: Does it make a difference in mathematical problem solving?. *Teachers College Record*, *119*(13).	11
Wilburne, J. M., & Dause, E. (2017). Teaching self-regulated learning strategies to low-achieving fourth-grade students to enhance their perseverance in mathematical problem solving. *Investigations in Mathematics Learning*, *9*(1), 38-52.	7

Excluded study	Exclusion criterion
Yarmohammadian, A., & Asli-Azad, M. (2012). Effects of metacognition training on the improvement of mathematical function in children with mathematic learning disability. *Advances in Cognitive Science*, *14*(1), 41-52.	2
Zohar, A., & Peled, B. (2008). The effects of explicit teaching of metastrategic knowledge on low- and high-achieving students. *Learning and Instruction*, *18*(4), 337-353.	7

Appendix B

Criteria and rationale for Weight of Evidence (WoE) ratings

WoE A: Methodological quality

WoE A is a generic judgment of whether a study is well executed (Gough, 2007). A published coding protocol designed to evaluate group experimental designs was used to make this judgment (Gersten et al., 2005). The protocol includes essential and desirable criteria. Adjustments were made to the wording of some questions in the protocol (see Appendix C for changes and rationale). The number of criteria required for WoE A ratings is provided in Table B1. The criteria for 'high' and 'medium' ratings were taken from Gersten et al. (2005), while the criteria for 'low' and 'very low' ratings were determined by the author because they were not provided by Gersten et al. (2005) but were desirable for the purposes of this review. A summary of scores from the coding protocol for each included study is provided in Table B2. Completed coding protocols for each included study are provided in Appendix D.

Table B1

Criteria for WoE A Ratings

WoE A rating	Criteria
3 High	Study meets at least 9 essential criteria *and* at least 5 desirable criteria
2 Medium	Study meets at least 9 essential criteria *and* at least 1 desirable criteria
1 Low	Study meets at least 6 but fewer than 9 essential criteria *or* does not meet any desirable criteria
0 Very low	Study meets fewer than 6 essential criteria

Table B2

Summary of Scores from the WoE A Coding Protocol

Criteria category	Chung & Tam (2005)	Fuchs et al. (2003)	Kajamies et al. (2010)	Pennequin et al. (2010)	Teong (2003)	Wang et al. (2019)	Zhu (2015)
Participant description (/3)	3	3	3	3	1	3	3
Intervention implementation (/3)	2	3	2	2	2	3	3
Outcome measures (/2)	2	2	1	2	1	2	1
Data analysis (/2)	0	2	2	2	1	2	2
Total essential (/10)	7	10	8	9	5	10	9
Total desirable (/10)	6	7	6	3	5	7	5
WoE A rating	1	3	1	2	0	3	3

Note. Criteria categories refer to essential criteria; desirable criteria were not categorised so only a total figure is reported. WoE A ratings are described as 'High' (3), 'Medium' (2), 'Low' (1), and 'Very Low' (0).

WoE B: Methodological relevance

WoE B is a judgment of the quality and relevance of the research design of a study to the review question (Gough, 2007). For this review, WoE B considered the relevance of the methodology for evaluating the effectiveness of metacognitive instruction at improving word problem-solving of children who were low-achievers in maths.

The coding protocol for WoE B was developed by the author and is provided in Table B3. There were 4 criteria which were categorised as 'participant allocation', 'comparison

intervention', 'outcome measures', and 'power analysis'. Scores were averaged across categories to produce a WoE B rating. A summary of scores from the coding protocol for each included study is provided in Table B4.

'Participant allocation' considered the extent to which participants were randomly allocated to experimental groups, whether this was at the individual level (high), classroom level (medium), or not random at all (low). Randomisation was valued because it reduces the possibility of selection bias (Barker et al., 2005).

'Comparison intervention' considered the extent to which the effect of metacognition instruction could be isolated. A no-intervention comparison group was considered weak because of potential Hawthorne effects (McCarney et al., 2007), where researcher attention was beneficial by itself regardless of intervention content. While an alternative intervention may have attenuated such effects, the ideal was judged to be a comparison group which received the same basic teaching within the same delivery parameters minus the explicit metacognitive components.

'Outcome measures' considered the extent to which pre- and post- problem-solving tests were comparable – in terms of difficulty, number of questions, number of steps – but also consisted of different questions to avoid practice effects. For a 'high' rating, the study needed to include a follow-up measure because this would provide data on whether participants could maintain any beneficial effects of intervention.

'Power analysis' considered the extent to which studies were adequately powered to detect effects. The software G*Power was used for these calculations (Faul et al., 2007). The statistical test used was 'ANOVA: Repeated measures, within-between interaction' because the primary measure of interest was the interaction of the between-measure of intervention/comparison group and the repeated-measure of pre/post score. The input parameters are provided in Table B5.

Table B3

Criteria and Rationale for WoE B Ratings

Criteria category		Criteria	Rationale
Participant allocation	3	Participants are randomly allocated to intervention and comparison groups	Randomisation reduces the possibility of selection bias.
	2	Participants are randomly allocated to intervention and comparison groups either at the individual level or as part of a block design at the classroom level	
	1	Participants are not randomly allocated to intervention and comparison groups or information is unavailable	
Comparison intervention	3	Comparison group receives the same teaching as the intervention group minus the metacognitive component with the same delivery parameters	The comparison group should isolate the effect of metacognition instruction as comprehensively as possible.
	2	Comparison group receives an alternative intervention or receives the same teaching as the intervention group minus the metacognitive component with different delivery parameters	
	1	Comparison group does not receive an alternative intervention	
Outcome measures	3	Parallel forms of outcome measures are taken at pre-test, post-test and follow-up	Pre-/post-measures should be comparable but avoid practice effects.
	2	Parallel forms of outcomes are taken at pre- and post-test but there is no follow-up	
	1	Outcome measures are taken at pre-test and post-test but they may not be parallel forms	
Power analysis	3	Sample size is adequate for all statistical analyses	Studies should be adequately powered to detect effects.
	2	Sample size may not be adequate for all statistical analyses	
	1	Sample size is inadequate for any analyses	

Table B4

Summary of Scores from the WoE B Coding Protocol

Criteria category	Chung & Tam (2005)	Fuchs et al. (2003)	Kajamies et al. (2010)	Pennequin et al. (2010)	Teong (2003)	Wang et al. (2019)	Zhu (2015)
Participant allocation	3	2	1	3	1	3	2
Comparison intervention	3	3	2	2	3	3	3
Outcome measures	1	2	3	2	3	2	2
Power analysis	1	3	1	3	2	2	3
WoE B rating	2	2.5	1.75	2.5	2.25	2.5	2.5

Note. WoE B ratings are described as 'High' for scores ≥ 2.5, 'Medium' for scores ≥ 1.5 and < 2.5, and 'Low' for scores < 1.5.

Table B5

*Input Parameters for Power Analysis Using G*Power*

Input parameter	Value
Effect size *f*	0.25 (medium)
α error probability	0.05
Power (1 – B error probability)	0.8
Correlation among repeated measures	0.5
Non-sphericity correction ε	1

WoE C: Topic relevance

WoE C is a judgment of the quality and relevance of the research evidence to the review question (Gough, 2007). The coding protocol for WoE C was developed by the author and is provided in Table B6. Scores were averaged across categories to produce a WoE C rating. A summary of scores from the coding protocol for each included study is provided in Table B7.

The key issues addressed were external validity, generalisability, and construct validity. External validity involves considering how effectively the interventions reflect what is possible and desirable in a standard educational setting.

Generalisability involves considering the extent to which a reader could determine whether the findings of a study would be applicable in another setting. The detail of participant characteristics is important for this and a breakdown is provided in Table B8. A further facet of generalisation is whether measurements were taken of participants' ability to generalise the knowledge taught to word problems in novel formats which they had not practised during intervention.

Construct validity involves considering how comprehensively metacognitive knowledge was taught in each intervention. To make this judgment, a further coding protocol was developed by the author, which is provided in Table B9. The metacognitive components were extracted from literature on the nature and extent of metacognitive knowledge (Flavell, 1979; Pintrich, 2002). In addition to the 11 components, a judgment was made of how much opportunity participants had to practise, because this is vital for metacognitive knowledge to be effectively assimilated (Livingston, 1996).

Table B6

Criteria and Rationale for WoE C Ratings

Criteria category		Criteria	Rationale
Participant characteristics and inclusion criteria	3	Inclusion criteria are described in terms of prior maths attainment based on a screening test administered by the researchers immediately prior to intervention. At least 7 relevant participant characteristics (see Table B8) are included.	Participants need to be described in sufficient detail for readers to determine the generalisability of findings to their own intended population.
	2	Inclusion criteria are described in terms of prior maths attainment but the measure may not have been administered by the researchers. At least 5 relevant participant characteristics are included.	
	1	Inclusion criteria in terms of prior maths attainment are very unclear or would be difficult to replicate. At least 3 relevant participant characteristics are included.	
	0	Inclusion criteria in terms of prior maths attainment are not provided. Fewer than 3 relevant participant characteristics are included.	
Intervention delivery	3	Intervention was delivered by existing educational staff in the regular educational setting	Findings will have greater external validity if it can be shown that schools can successfully deliver the intervention with existing staff in the regular setting.
	2	Intervention was delivered by research personnel in the regular educational setting	
	1	Intervention was delivered by research personnel outside the regular educational setting	
	0	Insufficient information is provided about the intervention delivery	
Setting generalisability	3	Participants were sampled from multiple mainstream or special schools	Most pupils who are low-achievers in maths attend mainstream or special schools. The findings will have greater external validity if they are replicated across multiple settings.
	2	Participants were sampled from a single mainstream or special school	
	1	Participants were sampled from alternative provision (e.g. pupil referral units)	
	0	Participants were not attending an education setting	

Criteria category		Criteria	Rationale
Metacognitive components	3	Intervention includes at least 1 component (see Table B9) of person, task, and strategy knowledge; at least 8 total components; and full opportunity to practise	Interventions should address the full range of metacognitive knowledge and provide sufficient opportunity to practise the teaching in order to maximise construct validity.
	2	Intervention includes at least 1 component of 2 of person, task, and strategy knowledge; at least 6 total components; and full opportunity to practise	
	1	Intervention includes at least 1 component of 2 of person, task, and strategy knowledge; at least 4 total components; and partial opportunity to practise	
	0	Intervention includes components of only 1 of person, task, and strategy knowledge; or fewer than 4 total components; or no opportunity to practise	
Generalisation of skills	3	The study measured task generalisation beyond the types of word problems practised during the intervention and took other measures of metacognitive attributes. The study included a follow-up measure.	Since metacognitive knowledge and skills are domain-general (Schraw, 1998), the extent to which participants were able to generalise and maintain their learning should be measured.
	2	The study measured either task generalisation or other metacognitive attributes. The study included a follow-up measure.	
	1	The study either included a measure of generalisation or a follow-up.	
	0	The study did not measure generalisation and did not take a follow-up.	

Table B7

Summary of Scores from the WoE C Coding Protocol

Criteria category	Chung & Tam (2005)	Fuchs et al. (2003)	Kajamies et al. (2010)	Pennequin et al. (2010)	Teong (2003)	Wang et al. (2019)	Zhu (2015)
Participant characteristics	2	1	3	1	1	3	2
Intervention delivery	2	3	2	0	0	2	3
Setting generalisability	2	3	3	2	2	3	2
Metacognitive components	1	1	2	2	1	0	1
Generalisation of skills	2	2	2	1	2	1	0
WoE C Rating	1.8	2	2.4	1.2	1.2	1.8	1.6

Note. WoE C ratings are described as 'High' for scores ≥ 2.5, 'Medium' for scores ≥ 1.5 and < 2.5, and 'Low' for scores < 1.5.

Table B8

Participant Characteristics Provided by Each Study

Participant characteristic	Chung & Tam (2005)	Fuchs et al. (2003)	Kajamies et al. (2010)	Pennequin et al. (2010)	Teong (2003)	Wang et al. (2019)	Zhu (2015)
Age or school year	✓	✓	✓	✓	✓	✓	✓
Gender	✓		✓	✓		✓	✓
Type of educational setting attended	✓	✓	✓	✓	✓	✓	✓
Type of developed environment of educational setting (i.e. urban, suburban, rural)		✓	✓			✓	✓
Ethnicity						✓	
Home language			✓				
Socio-economic background						✓	
Measures of general cognitive abilities	✓		✓			✓	✓
Measures of prior maths attainment	✓	✓	✓	✓	✓	✓	✓
Total	5	4	7	4	3	8	6

Table B9

Components of Metacognitive Knowledge Explicitly Taught as Part of Intervention

Metacognitive component	Chung & Tam (2005)	Fuchs et al. (2003)	Kajamies et al. (2010)	Pennequin et al. (2010)	Teong (2003)	Wang et al. (2019)	Zhu (2015)
Person knowledge (of oneself and others as cognitive processors)							
Strengths and weaknesses		✓				✓	
Motivation (e.g. self-efficacy, goals, value, interest)							
Articulating and sharing knowledge with others		✓	✓	✓			
Task knowledge (of what variations imply for managing cognitive enterprises)							
In which tasks different strategies are most appropriate		✓	✓	✓	✓		
How tasks of varying difficulty may require different cognitive strategies							
How local situational and general social, conventional, and cultural norms may influence the use of different strategies							
Strategy knowledge (of how cognitive enterprises can be effectively achieved)							
Planning (before attempting a task)	✓		✓	✓	✓	✓	✓
Monitoring (checking one's work during a task and adapting if necessary)	✓		✓	✓	✓	✓	✓
Evaluating (checking results after a task)	✓	✓	✓	✓	✓		✓
Information acquisition (e.g. elaborating, mnemonics, organising)	✓		✓	✓			✓
Problem-solving heuristics (e.g. means-ends analysis, working backwards from desired goal state)			✓				✓

Metacognitive component	Chung & Tam (2005)	Fuchs et al. (2003)	Kajamies et al. (2010)	Pennequin et al. (2010)	Teong (2003)	Wang et al. (2019)	Zhu (2015)
Opportunity to practise (the above facets of knowledge)							
Full – opportunity to practise all taught components	✓	✓	✓	✓	✓	✓	✓
Partial – opportunity to practise some taught components							
Total (excluding opportunity to practise)	4	4	7	6	4	3	5

Appendix C

Changes made to the WoE A coding protocol (Gersten et al., 2005)

Deletions are shown by ~~strikes through text~~, additions are in {curly brackets}, and rationale for the changes are in *[italics and square brackets]*. All other text was kept as original. As a result of these changes, there were ten desirable criteria rather than the original 8. The criteria of Gersten et al. (2005) for a 'high' quality study were altered to require 5, rather than 4, desirable criteria, which still constituted 50%.

Essential Quality Indicators

Quality indicators for describing participants

Was sufficient information provided to determine~~/confirm~~ whether the participants demonstrated the ~~disability(ies) or~~ difficulties presented?

[Rationale: to enhance readability and reflect the preference of 'difficulties' over 'disabilities' in relation to learning]

Were appropriate procedures used to increase the likelihood that relevant characteristics of participants in the sample were comparable across conditions?

Was sufficient information given characterizing the interventionists or teachers provided? Did it indicate whether they were comparable across conditions?

Quality indicators for implementation of the intervention and description of comparison conditions

Was the intervention clearly described ~~and specified~~?

[Rationale: to enhance readability]

Was the fidelity of implementation described and assessed?

Was the nature of services provided in comparison conditions described?

Quality indicators for outcome measures

Were multiple measures used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalised performance?

Were outcomes for capturing the intervention's effect measured at the appropriate times?

Quality indicators for data analysis

Were the data analysis techniques appropriately linked to key research questions and hypotheses? Were they appropriately linked to the unit of analysis in the study?

Did the research report include not only inferential statistics but also effect size calculations?

Desirable Quality Indicators

Was data available on attrition rates among intervention samples? ~~Was severe overall attrition documented? If so, is attrition comparable across samples? Is overall attrition less than 30%?~~

[Rationale: it was judged that 2 separate issues were being addressed by a single series of questions so they were separated (see below for further alterations)]

Was severe overall attrition {30% or more} ~~documented~~ {avoided}? If so, is attrition comparable across samples? ~~Is overall attrition less than 30%?~~

[Rationale: to enhance readability and to make this a positively-worded question so that a 'Yes' answer would count as a positive point]

Did the study provide not only internal consistency reliability but also test-retest reliability and interrater reliability (when appropriate) for outcome measures? ~~Were data collectors and/or scorers blind to study conditions and equally (un)familiar to examinees across study conditions?~~

[Rationale: it was judged that 2 separate issues were being addressed by a single series of questions so they were separated]

Were data collectors and/or scorers blind to study conditions and equally (un)familiar to examinees across study conditions?

Were outcomes for capturing the intervention's effect measured beyond an immediate post-test?

Was evidence of the criterion-related validity and construct validity of the measures provided?

Did the research team assess not only surface features of fidelity implementation (e.g. number of minutes allocated to the intervention or teacher/interventionist following procedures specified), but also examine quality of implementation?

Was any documentation of the nature of instruction or series provided in comparison conditions?

Did the research report include actual audio or videotape excerpts {or examples of paperwork} that capture the nature of the intervention?

[Rationale: several studies included appendices with paperwork that captured the nature of the intervention, such as worksheets and lesson plans. It was deemed that this provided similar evidence of ecological validity to audio or videotape excerpts]

Were results presented in a clear, coherent fashion?

Appendix D

Example of a complete WoE A coding protocol (Gersten et al., 2005)

Study: Chung, K. K., & Tam, Y. H. (2005). Effects of cognitive-based instruction on mathematical problem solving by learners with mild intellectual disabilities. *Journal of Intellectual and Developmental Disability*, *30*(4), 207-216.

Essential Quality Indicators

Quality indicators for describing participants

Was sufficient information provided to determine whether the participants demonstrated the difficulties presented?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Were appropriate procedures used to increase the likelihood that relevant characteristics of participants in the sample were comparable across conditions?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Was sufficient information given characterizing the interventionists or teachers provided? Did it indicate whether they were comparable across conditions?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Quality indicators for implementation of the intervention and description of comparison conditions

Was the intervention clearly described?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Was the fidelity of implementation described and assessed?

☐ Yes

☒ No

☐ Unknown/Unable to Code

Was the nature of services provided in comparison conditions described?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Quality indicators for outcome measures

Were multiple measures used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalised performance?

☒ Yes – partially; word problems were classified as either 'similar' to those practised during the intervention or 'transfer' which 'were used to examine whether the use of a particular format could enable the students to solve a wider range of problems'

☐ No

☐ Unknown/Unable to Code

Were outcomes for capturing the intervention's effect measured at the appropriate times?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Quality indicators for data analysis

Were the data analysis techniques appropriately linked to key research questions and hypotheses? Were they appropriately linked to the unit of analysis in the study?

☐ Yes

☒ No – The study did not include pre-test scores in the final analysis so, although the pre-test scores of all groups were comparable, there is no statistical data reported on whether the groups made pre- to post-test improvements

☐ Unknown/Unable to Code

Did the research report include not only inferential statistics but also effect size calculations?

☐ Yes

☒ No

☐ Unknown/Unable to Code

Desirable Quality Indicators

Was data available on attrition rates among intervention samples?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Was severe overall attrition (30% or more) avoided? Is attrition comparable across samples?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Did the study provide not only internal consistency reliability but also test-retest reliability and interrater reliability (when appropriate) for outcome measures?

☐ Yes

☒ No – only interrater reliability reported (coefficient .85)

☐ Unknown/Unable to Code

Were data collectors and/or scorers blind to study conditions and equally (un)familiar to examinees across study conditions?

☐ Yes

☒ No

☐ Unknown/Unable to Code

Were outcomes for capturing the intervention's effect measured beyond an immediate post-test?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Was evidence of the criterion-related validity and construct validity of the measures provided?

☐ Yes

☒ No

☐ Unknown/Unable to Code

Did the research team assess not only surface features of fidelity implementation (e.g. number of minutes allocated to the intervention or teacher/interventionist following procedures specified), but also examine quality of implementation?

☐ Yes

☒ No

☐ Unknown/Unable to Code

Was any documentation of the nature of instruction or series provided in comparison conditions?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Did the research report include actual audio or videotape excerpts or examples of paperwork that capture the nature of the intervention?

☒ Yes – examples of materials used in all three groups

☐ No

☐ Unknown/Unable to Code

Were results presented in a clear, coherent fashion?

☒ Yes

☐ No

☐ Unknown/Unable to Code