Cognitive Science in the Classroom:

Evidence and Practice Review

July 2021

Thomas Perry[1]
Rosanna Lea[1]
Clara Rübner Jørgensen[1]
Philippa Cordingley[2]
Kimron Shapiro[1]
Deborah Youdell[1]

with
Julia Harrington[3]
Amy Fancourt[3]
Paul Crisp[2]
Niall Gamble[1]
Christina Pomareda[1]

[1] **University of Birmingham, England**
[2] **Centre for the Use of Research and Evidence in Education (CUREE), England**
[3] **Queen Anne's School and BrainCanDo, England**

Date of systematic review searches: *August 2020*

Date of practice review data collections: *November to December 2020*

Please cite this report as follows:

Perry, T., Lea, R., Jørgensen, C. R., Cordingley, P., Shapiro, K., & Youdell, D. (2021). Cognitive Science in the Classroom. London: Education Endowment Foundation (EEF). The report is available from: https://educationendowmentfoundation.org.uk/evidence-summaries/evidence-reviews/cognitive-science-approaches-in-the-classroom/

A summary report for this review is available from: https://educationendowmentfoundation.org.uk/evidence-summaries/evidence-reviews/cognitive-science-approaches-in-the-classroom/

Acknowledgements

We would like to thank the Education Endowment Foundation for commissioning this review. Particular thanks go to Jonathan Kay and Harry Madgwick for support and advice during the project, especially during the opening and closing stages of the review amidst the uncertainty and challenges posed by the national pandemic lockdowns.

This project has benefitted greatly from the advice and constructive challenge from the advisory group. We would like to thank all members of the group and their generous contributions in and around the group meetings.

The project advisory group members were as follows:

- Dr Robin Bevan
- Prof. Robert Coe
- Dr Iroise Dumontheil
- Dr Amy Fancourt
- Dr Davinia Fernández-Espejo
- Julia Harrington
- Dr Niki Kaiser
- Mark Stow
- Prof. Hillevi Lenz Taguchi
- Sonia Thompson
- Prof. Sam Twiselton

We would also like to thank Professor Steve Higgins whose advice helped us clarify our approach as we were designing our searching, coding, and analytical strategies when creating the review protocol.

Finally, we would like to thank all the teachers and school leaders who have taken part in our interviews and questionnaire. Your perspectives have sharpened our focus and strengthened our discussion of the results and their implications. We have drawn on your collective wisdom and sought to connect your contributions with the evidence to describe, explain, and pose valuable questions about the theory and practice of cognitive science in the classroom.

A warm thank you to you all.

Table of Contents

Part A: Review methods and background information

A1. Background and review rationale

Scientific, policy, and practical background

Cognitive science, policy, and practice

Learning sciences form an interdisciplinary field that draws on cognitive science, educational psychology, computer science, anthropology, sociology, information sciences, neurosciences, education, instructional design, and numerous other areas (Sawyer, 2006). The central role of cognition in learning, and the brain's role in processing and storing information, arguably places cognitive science at the heart of the learning sciences. Thus, many influential publications aimed at practitioners focus on cognitive science.

Two areas of cognitive science have been especially influential in education:

- **cognitive psychology**—concerned with mental processes including perception, thinking memory, attention, and learning; cognitive psychology is underpinned by interpretive, behavioural, and observational methods; and
- **cognitive neuroscience**—concerned with the brain and the biological processes that underlie cognition; cognitive neuroscience is underpinned by brain imaging technologies such as electrophysiology (EEG) and functional imaging (fMRI).

For many decades, the dominant science for informing education practice has been cognitive psychology. Multiple publications directed at educators and a lay public aim to make accessible lessons for learning drawn from cognitive and educational psychology (for recent examples, see Weinstein, Sumeracki and Caviglioli, 2018; Kirschner and Hendrick, 2020; also see Deans for Impact, 2015; Pashler et al., 2007).

More recently, neuroscience research has gathered a great deal of momentum, not least because of the advent of new imaging technologies that enable much finer-grained research; strong claims are made for the application of educational neuroscience (Goswami, 2006; Howard-Jones, 2014) or 'neuro-education' (Arwood and Meridith, 2017). Again, there is a body of publications focused on the implications of the science for classroom practice (for example, Tibke, 2019; Jensen and McConchie, 2020).

Practice-focused accounts of cognitive science have proved highly influential for educators looking for a scientific basis to inform and improve their practice.

Both areas of cognitive science are currently and increasingly informing interventions, practice, and policy in education. Of particular interest to education has been basic cognitive psychology and cognitive neuroscience research in brain structure and function, motivation and reward, short-term (or working) and long-term memory, and cognitive load. Cognitive science forms a key part of the evidence underpinning the Ofsted Education Inspection Framework (Ofsted, 2019). Ofsted reports 'moderate to strong evidence of practices that can be used to enhance learning across phases and remits'; it refers to strategies including spaced practice, interleaving, retrieval practice, elaboration, and dual coding and discusses the 'important contribution' of cognitive load theory (Ofsted, 2019, p.19). This commitment to cognitive science is also evident across the National Professional Qualifications frameworks and across the current Teaching Standards for early career teachers and

the associated early career teacher training programmes; there are clear expectations that newly qualified teachers will be well informed about cognitive science—particularly concerning cognition and information storage and retrieval (memory)—and that they will develop skills to apply this knowledge in the classroom.

Other areas of cognitive science have potential implications for education but are yet to enter widespread educational attention or practice. Not yet of interest to policymakers but with translational potential, for example, is research in cognitive neuroscience exploring synchrony of brain oscillations and memory formation (Clouter, Shapiro et al., 2017). 'Brain oscillations' refers to the electrical activity generated by the brain in response to stimuli. There is growing interest in how these oscillations can synchronise between individuals during social interactions ('brain-to-brain synchrony'). The significance of brain-to-brain synchrony has recently been tested in education interactions where it has been found in pair and group interactions, is shown to predict memory retention, and has been connected to forms of instruction, teacher-student relations, and learning outcomes (Bevilacqua et al., 2018; Davidesco et al., 2019; Dikker et al., 2017).

Learning concepts derived from cognitive science vary in the extent to which they enjoy consensus in the scientific community regarding the science and its educational implications.

The basic research into cognitive load, for example, is well-established, drawing on behavioural psychology and a conceptual amalgamation of cognitive load and attention applied to educational principles of learning (for overviews, see de Jong, 2010; Sweller, 2016). It is yet, however, to be tested by the latest imaging neuroscience techniques. Furthermore, whether understanding of cognitive load in brain function and activity over very short (sub-second) units of time can be mobilised effectively to inform planning and delivery of teaching and learning that takes place over substantially larger units of time, and in a way that will have a measurable beneficial effect, remains to be demonstrated.

Practice informed by cognitive science

Cognitive science is gaining increasing influence in education and a multitude of existing and developing classroom practices are currently described as being 'informed' or 'inspired' by cognitive science. Some warn against the over- or mis-interpretation of cognitive science evidence for use in education (Alferink and Farmer-Dougan, 2010) and the risk of 'neuromyths' (Howard-Jones, 2014, 2018; Purdy, 2008), especially for commercial educational products claiming a basis in neuroscience.

Insights from cognitive science have the potential to both displace, complement, and add to complex and varied understandings of effective classroom practice.

Within schools, many techniques that are currently described as inspired by cognitive science may have been practised previously using a different rationale, including ones drawing on cognitive science. As discussed in our first project advisory group meeting (see Appendix 2 for a summary), techniques currently being described as inspired by cognitive science may have been practised previously by 'hunch' (for example, quizzes) rather than being explicitly informed by cognitive science. Many of these resonate with established understandings of effective pedagogy and some may have been practised by teachers previously with no specific reference to cognitive science (Alferink and Farmer-Dougan, 2010; Willis, 2009). Cognitive science, as well as potentially identifying new or improved practices, can also provide a shared understanding and common language about existing techniques and help establish why some commonly used teaching methods work or do not.

Like all learning theories, ideas from cognitive science must be applied to specific subjects, phases, students, and learning contexts. Another point emerging from the first advisory group meeting was that for the review to realise its potential benefit it must connect cognitive science informed teaching and learning strategies to contexts and subjects and consider—where possible—the influence of these on the results. To some extent, mediating and moderating factors are reported in classroom trial evidence and therefore amenable to analysis within this review. Scoping work (see below), however, suggested that the classroom trial literature is yet to reach a point of maturity to enable impact evidence to be comprehensively assessed across subjects and contexts. Moreover, as cognitive science informed practice is, in many areas (for example, subjects, phases, and practitioner groups), still in a nascent state, going beyond trial-based evidence by collecting empirical evidence and reviewing practice-focused documents—as per our practice review sub-strand—is valuable to lay the groundwork for future practitioner-facing accounts of cognitive science as well as better understand its relation to pre-existing practice.

Basic science and applied science

One of the most powerful reasons for looking to cognitive science to inform education is that it seeks to offer robust evidence that reveals something fundamental about memory, learning, and the brain. This growing understanding will not necessarily or automatically yield valuable insights for classroom practice; the extent to which findings from controlled settings such as laboratories are applicable in real classrooms remains to be seen. On the one hand, understanding the fundamentals of learning will (or so the argument goes) be highly applicable across contexts (that is, have high external validity) as humans share the same basic cognitive architecture and utilise the same cognitive processes during learning experiences. On the other hand, the context in which the basic scientific evidence is produced—often controlled settings such as the laboratory—can be far removed from the classroom teaching and learning context it looks to inform ; it has, in other words, low 'ecological validity'—validity in real classrooms, across the curriculum, and for different pupil groups. Evidence from basic science can (a) require a greater degree of translation to get the strategy working in practice and (b) potentially be reductive through experimental control when isolating principles and the effects of specific strategies. Put simply, teaching and learning 'set pieces' delivered by researchers, designed to focus on one cognitive process, may not work well in the hands of real teachers negotiating the demanding and complex contexts and problems of real classroom environments.

In our view, there is value in considering both applied and basic science side-by-side when looking to support and inform classroom practice. We share the enthusiasm of many teachers and researchers in the potential insights and understanding from basic science. We do not, however, assume that basic science necessarily or easily translates into effective classroom practice. In short, something that works in the laboratory may, or may not, work well in the classroom. A vital purpose of this review is to make this distinction and examine the evidence from the classroom.

The focus of this review is a systematic review of the applied *science and, in particular, evidence from ecologically-valid classroom trials of strategies that are informed by cognitive science. We want to know whether cognitive science techniques work in real classrooms, across the curriculum, and for different pupil groups.*

The Education Endowment Foundation (EEF) is at the forefront of promoting evidence-informed practice in education in England and this review is part of its work summarizing the best available evidence for teaching and learning to support teachers and leaders to raise the attainment of 3- to 18-year-olds, particularly those facing disadvantage. We return to this point about types of evidence and

the current state of evidence for cognitive science in the classroom in connection with the review problem statement, below, after briefly discussing relevant narrative and systematic reviews in this area.

The other key idea touched on here is that of 'translation' of science for policy and practice. The challenges of translation further emphasise the importance of reviewing applied evidence; it also frames our discussion and questions sections, which examine the current state of knowledge about the principles and practices of applying lessons from cognitive science in the classroom.

Our review is founded on the view that translation of evidence from basic science is neither simple nor unproblematic.

There remains much to be understood concerning the translation of cognitive science evidence to education and its application in the classroom, as well as with regard to the underpinning basic science itself (Fischer et al., 2010; Aronsson and Lenz Taguchi, 2017; Rose and Rose, 2013). Where cognitive science has been translated and applied within classroom interventions and techniques, it cannot safely be assumed that this will have the expected impact on pupils' learning, however strong the basic science evidence-base. Selected translational work reviewed when scoping this review revealed a mixed picture. For instance, classroom-based manipulation of reward based on neuroscientific work on memory has shown null results in an EEF trial (Mason et al., 2017) and neuroscience-based language learning interventions in the classroom—which have elsewhere been effective (Goswami, 2015; Kyle et al., 2013)—have recently shown no significant benefit (Worth et al., 2018). These results do not necessarily (but may) bring the underpinning science and its evidence-base into question. They do, however, highlight the value of seeking evidence with both high internal and ecological validity as a basis for practice and distinguishing and evaluating the evidence-base in these terms.

Existing narrative and systematic reviews in this area

Teaching and learning strategies informed by cognitive science

The previous section cites numerous practice-focused books and reports that provide narrative reviews relating to cognitive science in the classroom. These sources were incorporated into this review by informing the review protocol—especially the conceptual framework, research questions, and search terms—and by synthesising and summarising their characterisation of the cognitive science literature and its practical implications in our discussion sections.

Our scoping searches for published work relating to the educational applications of cognitive science[1] yielded only three systematic reviews/meta-analyses published since 2000. These pieces focused on, first, RCTs in education research (Connolly, Keenan, and Urbanska, 2018), second, teacher-led neuroscience-based RCTs (Churches et al., 2020), and third, mathematics interventions (Kroeger, Douglas-Brown, and O'Brien, 2012).

There are also several recent and influential non-systematic reviews that have influenced the focus and design of the present review; below, we briefly summarise selected literature we reviewed during scoping that impacted the aims and design of this review. The Connolly et al. and Churches et al. systematic reviews are described immediately below; we make references to Kroeger et al. towards the end of this section; all other pieces mentioned are *not* systematic reviews.

[1] These searched for cognitive science terms in general rather than specific strategies.

Connolly et al. (2018) produced the first systematic review to evaluate all RCTs conducted in education from 1980 to 2016. While the study did not examine the evidence of effects for cognitive-science-inspired strategies specifically, the yield of 1,017 unique randomised controlled trials (RCTs) disputes the claims that it is impossible to undertake quality RCTs in education. However, as a rapidly growing field, there is now a sufficient threshold to conduct a systematic review of strategies that specifically focus on memory and information acquisition and retrieval, which, as alluded to above, is an area of substantial interest to science, policy, and practice. Also instructive for this review were the methodological recommendations made by Connolly et al. who explain that to connect to practice meaningfully, we need 'more nuanced and sophisticated trials' that 'are acutely aware of the contingent and context-specific nature of educational interventions' (Connolly et al., 2018, p.14). This recommendation connected with discussions about our focus during protocol development (see below).

The recent meta-analysis conducted by Churches et al. (2020) analysed the findings of 34 teacher-led RCTs to bridge the gap between neuroscience and educational practice. In essence, teachers designed and implemented cognitive science informed interventions, with each trial focusing on a single strategy (for example, attention, retrieval practice, spaced practice, or interleaving). Overall, this important work demonstrated that cognitive science informed strategies can indeed benefit pupil outcomes. Given its relevance, we have included an analysis of this study in the main results.

Other reviews are also of note with respect to the utility of specific cognitive science informed strategies (Dunlosky et al., 2013; Weinstein et al., 2018). A 2013 monograph (Dunlosky et al., 2013) reviewed the literature to ascertain the relative utility of ten learning techniques, with some consisting of the 'core' strategies on which we focus (for example, interleaving and spaced practice), but also including ones that students commonly self-report using (for example, highlighting and rereading). Similarly, Weinstein et al.'s 'Teaching the Science of Learning' (2018) highlights six cognitive strategies: spaced practice, interleaving, retrieval practice, elaboration, concrete examples, and dual coding. Such reviews (also see scoping bibliography, Appendix 2) provide valuable and accessible overviews of the most popular techniques; however, these reviews are generally not systematic and the scientific bases are not explored in-depth—one of the objectives of the present review. Such studies helped this review confirm its focus and selection of strategies for review. Also of note was information provided on principles for how, for what, and in what conditions the strategies are likely to work. Dunlosky et al. (2013), for instance, considered the role of several mediators and moderators such as state and trait-based student characteristics (such as age and intelligence), learning conditions (for example, group versus individual learning), and task type. Our review has been conscious of these moderators that may influence the effectiveness of cognitive science informed interventions and we designed our approach to screening, data extraction, and analysis to keep them in view.

The EEF has already commissioned several reviews with relevance to the application of cognitive science to the classroom and it is important to distinguish how the present review builds upon, but remains distinct from, these. Of particular note is the 2014 EEF review of educational interventions and approaches informed by neuroscience (Howard-Jones, 2014): this provides a helpful overview of the neuroscientific approaches of relevance to the classroom and an appraisal of the evidence-base for these as it stood in 2014. The present review does not aim to replicate this review but to re-examine cognitive science informed approaches and the studies underpinning them within the context of a systematic review, integrating the evidence from those studies with more recent ones and supporting the development of the EEF Education Database (see objectives below). A systematic approach—along with review sub-strands focused on the underpinning science and practice—is needed to gain a more current, holistic, and coherent picture of a rapidly growing field.

There have also been reviews concerning specific classroom strategies inspired by cognitive science and specific subject areas. For example, reviews of game-based learning (Jabbar and Felicia, 2015), spaced repetition (Kang, 2016), and the role of instructional explanations in example-based learning (Wittwer and Renkl, 2010). Another notable example is Kroeger, Douglas-Brown, and O'Brien's (2012) review of neuroscience-informed mathematics intervention programmes identifying three eligible programmes.

Even where systematic reviews exist for particular strategies, an overarching systematic review is needed to update the evidence and simultaneously capture, analyse, and evaluate all relevant teaching and learning strategies concerning both their underpinning cognitive science *and* their applicability to classroom contexts by restricting studies to those with a classroom trial design.

In summary, to our knowledge, there are no reviews that systematically review evidence of impact on pupils from classroom interventions, that are recent, and that have broad coverage of cognitive science informed strategies for *acquiring and retaining knowledge.*

Wider Cognitive Science-Informed Strategies

The current review takes as its focus cognitive science informed teaching and learning strategies for acquiring and retaining knowledge. Cognitive science, however, is a broad and interdisciplinary field that (as reflected in the 2014 EEF review) includes research on the physical, emotional, and social conditions that support the processing, acquisition, and use of knowledge.

These wider considerations fall at the boundary of this review. We recognise that clearly delineating cognitive science informed strategies for acquiring and retaining knowledge from other cognitive science informed classroom strategies is not always clear-cut. How this boundary is navigated in terms of review methodology (for example, inclusion and exclusion criteria, data extraction, analysis, and the sub-review strands) is set out in detail below and the next section is devoted to defining and justifying the review focus in more detail. Here, we briefly outline reviews that sit on this boundary and thereby inform and define the present review.

First are cognitive science topics relating to **social and emotional aspects of cognition**. The EEF has already published an evidence review on social and emotional learning (SEL) strategies (Wigelsworth et al., 2020), elements of which (for example, mindfulness and stress reduction) are present in the cognitive science literature and relevant to the present review. School-based SEL strategies aim to help pupils acquire and effectively apply the knowledge, attitudes, and skills necessary to understand and manage emotions, set and achieve positive goals, establish and maintain positive relationships, and make responsible decisions.[2] Such trials demonstrate positive effects on wellbeing, behaviour, and the more social aspects of school life; some studies also suggest that emotional competence predicts academic achievement (for a meta-analysis, see MacCann et al., 2020), indicating potential cognitive benefits.

A second key area adjacent to this review is that of **metacognition**. Of particular note, an EEF review on metacognition and self-regulated learning (Muijs and Bokhove, 2020), with an accompanying Toolkit strand, explored the evidence-base regarding teaching and learning strategies related to pupils' metacognition (in which pupils are encouraged to think about their own learning explicitly). To avoid the duplication of research efforts, the present review does *not* include metacognition as one of its strategies, using this EEF review to define its boundaries. Finally, we were aware that there was a concurrent EEF systematic review taking place focusing on **feedback**. Similarly, the present review

[2] Also see http://casel.org

does not include feedback processes as cognitive science informed classroom strategies unless there are specific links to the core cognitive science concepts and strategies identified below.

In overview, we have scoped studies in these boundary areas to: (a) avoid duplication of effort or content within the EEF's evidence and resource base, (b) identify contextual and moderating factors for cognitive science informed teaching and learning strategies for acquiring and retaining knowledge, and (c) identify boundary cases that meet the review inclusion criteria and thereby avoid an overly reductive account of strategies within the core focus.

A2. Focus and problem statement

Problem statement

As described above, over the last decade—and in particular over the last few years—there has been growing interest in the practical application of findings from cognitive science to classroom practice. There are now a substantial number of recent and influential reviews and impacts on practice and policy. In large part, reviews identify and advance implications for classroom practice by considering laboratory-based research and plausible interpretations of the basic science. In the previous section, we discussed the importance of distinguishing *basic* from *applied* cognitive science and briefly outlined their strengths and limitations, and noted the value of considering them jointly.

In summary, several challenges are apparent with the current state of applied cognitive science literature and practice:

- many techniques require considerable translation for application to the classroom;
- the research and practice of translation is emergent and variegated, with particularly rapid change in the last few years;
- it cannot be taken for granted that techniques with firm foundations in the basic science will be, or are being, successfully applied in effective interventions and practices in the classroom: initial evidence demonstrating successful application is mixed;
- the extent to which emerging practice has fidelity to the underlying cognitive science is unclear; and
- education research offers incomplete understandings of the influences on, and mechanisms of, learning (Youdell and Lindley, 2018); how cognitive science informed practice relates to this research and existing methods, including other evidence-informed approaches, is unclear.

These issues bring into focus the value of applying and testing cognitive science principles in realistic classroom settings and of exploring the connections between cognitive science theory and practice. Our outline of previous relevant reviews above revealed a growing general literature of trials in education. With increasing application of the basic science we would expect increasing numbers of trials within this literature to be of strategies and interventions informed by cognitive science. However, this literature has yet to be subjected to a systematic impact review; any reviews that have taken place have focused on specific strategies or have not used rigorous systematic review methods.

We hold that without this applied evidence, practice in this area can only be said to be evidence-informed in a weak sense (that is, having only limited consonance with basic cognitive science). What is required as a basis for a stronger form of evidence-informed practice is a systematic review of classroom interventions, appraised in relation to both internal and external, 'ecological' validity.

This is the ideal time for this systematic review of cognitive science in the classroom. Without a systematic review of the applied evidence, the current and potential impact on pupil outcomes of various cognitive science informed interventions and techniques remains uncertain.

This uncertainty and incompleteness contrasts with a widespread positive view about the potential benefits of applied cognitive science, including education policymakers' keen attention to advances in cognitive psychology and neuroscience and efforts to use these to inform teacher education, guidance for classroom practice, and school inspection.

These gaps in understanding discussed above bring us to the three strands of this review:

- Our core strand—and central focus—concerns the evidence of impact in the classroom of cognitive science informed interventions and techniques.

The sub-strands examine:

- first, basic cognitive science theories and concepts, as described in the scientific and practitioner-guidance literatures; and
- second, what cognitive science informed practice and interventions look like in the classroom and how they vary according to the subject, context, and pupil characteristics.

Below, we provide an overview of, and background information on, these strands and the links between them. Before doing so, we set out our core definitions and conceptual frameworks and return to the question of the focus and boundaries of the review.

Definitional and conceptual frameworks

Translation, implementation, and knowledge transfer

Overarching concepts informing the design of this review include translation, implementation, and knowledge transfer.

We hold that there are rarely singular, unambiguous implications of findings from cognitive science for the classroom; rather, a process has necessarily taken place whereby the basic science has been interpreted and translated into education interventions and techniques and operationalized and implemented across subjects and learning contexts.

Important considerations, therefore, when assessing the implications of the results of the trial-based evidence include the translation of the cognitive science, the ecological validity of the intervention, and the fidelity of implementation to the intervention design (Youdell, Lindley, Shapiro, Sun and Leng, 2020). This review has been designed to include consideration of translation and implementation within the core review and sub-strands. This allows us to examine, for example, whether interventions are likely to work in other settings or whether inconsistent or null findings are likely to stem from weaknesses in the underpinning science, its translation, or implementation. While the core focus remains on the evidence of impact (see below), these wider considerations enable the results to be contextualised, support the process of drawing implications for research and practice from the core findings, and inform further research and practice in applied cognitive science.

Table A2.1 summarises our theory of translation from cognitive science to classroom practice to the production of an evidence-base on impact. Below, we indicate the scope of the review strands in relation to this overarching theory.

Figure A2.1: Theory of translation

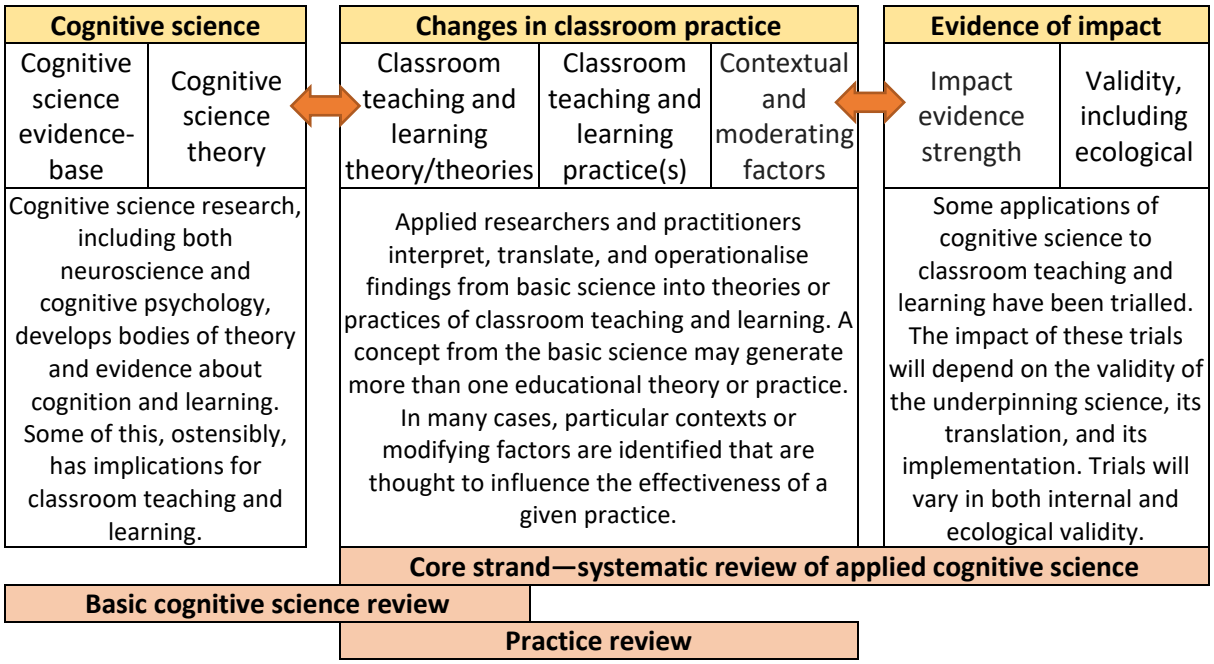


Table A2.1 also serves as a summary of the overall landscape investigated within this review. Our focus is the core strand, where we systematically review evidence of impact (right). We also connect this to the theory and evidence of cognitive science informed classroom practice (centre). In other words, we discuss *why* and *how* as well as *if* cognitive science informed strategies work. The sub-strands of the review are also evident: first, the practice review has been designed to explore in greater depth the theory and practice of the cognitive science informed strategies (middle), beyond what we can currently conclude from the applied evidence (right); similarly, we have reviewed the principles and findings of cognition and learning from the basic science (left) and connect these with our discussion of the cognitive science strategies in focus, again, setting out prominent ideas even where these cannot be connected to the applied evidence. The applied cognitive science was likely to be patchy, uncertain, and quite general in what conclusions might be reached. In contrast, practice is far more granular and comprehensive and basic science more certain and coherent in its principles. In this sense, there are large gaps between theory, practice, and evidence. Through our sub-strands, and discussion sections, we are better placed to make sense of the applied evidence (right), make connections to practice and theory, and identify the next steps for the development of the evidence-base.

The present review encompasses the whole of the process from cognitive science (Which strategies show potential from perspective of basic science?), to impact assessment (Which practices have been subjected to classroom trials and what does this evidence reveal about their effectiveness?), to practice (Which strategies do teachers actually *use?). Our review also aims to connect the core evidence-base to a wider range of strategies through its sub-strands, including those that may be commonly used in educational contexts but perhaps without an overt basis in cognitive science.*

Finally, it is also valuable to note that cognitive science informed practice can be supported by knowledge transfer in both directions (science-practice and practice-science): cognitive science techniques already part of established practice can be translated (or traced) back to the cognitive

science; basic science may shed light on why, and in which contexts, some commonly used practices work and others do not; and details of practice and applied research might identify a wider range of relevant operative and contextual factors and avenues for future basic research.

Conceptual framework

The review strands are held together by our conceptual framework, which was tested and refined throughout the systematic review process (Gough, Oliver and Thomas, 2017). The conceptual map (see Appendix 2) presents the initial concepts of interest. These have been broadly divided into categories:

- approaches informing the **design of classroom teaching and learning** (core)—spaced practice, interleaving, dual coding, retrieval practice, and cognitive load;
- approaches that involve **physical factors** (wider)—exercise, nutrition and hydration, and sleep;
- approaches that involve the **motivational or emotional state of the learner** (wider)—mindfulness, stress and anxiety reduction, social and emotional learning, and reward or game-based learning; and
- approaches that involve **direct manipulation or measurement of neural activity** (wider)—transcranial electrical stimulation and brain-to-brain synchrony.

This evolving conceptual map: (a) helped us specify our focus in terms of the theories from cognitive science (including related concepts or synonyms), (b) provided definitions for, and an overview of, each theory, related practices, and variations in these, (c) identified relationships and boundaries between them, and (d) identified known modifiers or contextual factors thought to influence the effectiveness or applicability of the theory in practice. It is important to acknowledge that these concepts overlap. A preliminary outline of some of the essential conceptual connections is indicated in the conceptual map.

The focus and scope of the review

The systematic review investigated approaches to teaching and learning informed by cognitive science that are commonly used in the classroom, with a particular focus on acquiring and retaining knowledge. This focus reflects the areas of cognitive science which have to date been the most influential for classroom practice and ostensibly have the most general application across the education sector.

While there are numerous formulations and accounts of crucial techniques informed by cognitive science relating to acquiring and retaining knowledge, the following concepts are the most widely known and were included in the EEF's initial review specification:

- spaced practice;
- interleaving;
- retrieval practice;
- dual coding; and
- (strategies to manage) cognitive load.

These concepts are briefly outlined in our conceptual map, along with examples of how these might be used in the classroom, as informed by the review scoping work (Appendix 2). Our core focus also included a wider and related set of strategies (see also Appendix 2). Indicative examples include the

use of concrete examples, highlighting during study (Dunlosky et al., 2013), elaboration and interrogation techniques, and techniques from 'brain training'. Beyond the core strategy list above, inclusion of the (sub)strategies beyond those for which we conducted targeted searches was based on the specified inclusion and exclusion criteria set out in the review protocol rather than through *a priori* specification of concepts for inclusion.

We conducted (a) targeted searches for core concepts, as described, and (b) more general searches for related and broader cognitive science strategies. We aimed to locate literature to review a broad and comprehensive set of cognitive science concepts focused on acquiring and retaining knowledge. *Clear and practicable categorisation of concepts and account of the relationships between them were developed alongside coding work undertaken during the review as the evidence emerged.*

The approach outlined is a form of framework synthesis where an *a priori* initial conceptual framework is modified during systematic review to organise the evidence and an outcome of the review (Gough, Oliver and Thomas, 2017). The tentative, working concept map produced from early scoping work (see previous) was the starting point for this framework development. The first section on the conceptual map, 'approaches informing the design of teaching and learning techniques', captures our core focus on knowledge acquisition and retention strategies. We have also identified several other areas relating to physical and emotional areas of cognitive science and areas involving direct measurement or manipulation of neural activity. As discussed below, these are at the boundary of our review. In the methods appendices (Appendices 1 to 3) we describe the rationale for inclusion and exclusion of studies across the review.

One clear message from the first advisory group (see Appendix 1) was that for the review to realise its potential benefits, there was a need to avoid a reductive view, not only of cognitive science, but also of the concept of learning and, consequently, of links between the two. There are several issues connected to successfully navigating this boundary.

- First, focusing exclusively on the five core concepts (as above) **risks an overly narrow focus** on the concepts currently popularised in policy and practice in a particular location (England). In effect, this sets the bounds of the review around cognitive science strategies—and in particular those from cognitive psychology—that are included in previous influential (but typically not systematic) reviews as opposed to a defined (conceptual and practical) set of inclusion and exclusion criteria.
- Second, a focus on 'cognitive science informed strategies for acquiring and retaining knowledge' might conceivably include **classroom strategies relating to social and emotional aspects of learning**, or physical or social factors supporting cognition. Previous research (for example, Howard-Jones, 2014) has, for example, highlighted that maths anxiety leads to greater activity in the amygdala region of the brain and reduced working memory. If, say, a simple classroom strategy was identified that organised a sequence of maths problems in a way that reduced anxiety, arguably this would promote the acquisition and retention of knowledge, be based on a strategy informed by cognitive science, and fall within the scope. Similarly, there may be interventions that employ multiple strategies that cut across the broad groups identified. One example of this might be motivational or emotional strategies that use quizzes (that is, retrieval practice) or concrete examples to engage learners.
- Third, and related to the previous point, many concepts identified within the wider boundary concepts of the working concept map may be **contextual, moderating, or mediating factors** for core strategies. For example, whether a 'low-stakes' quiz used for retrieval practice *feels* low-

stakes (and does not induce stress or anxiety) may depend on the emotional strategies employed by the teacher to promote cognition. The boundaries between 'active ingredients' from cognitive science and the contextual or moderating factors may not be entirely clear. Potential overlaps and links with other concepts are highlighted within our concept map (see Appendix 2); many of these overlaps are between concepts from the core focus area and the wider areas in the map.

- Fourth, **wider concepts that meet the review criteria may not be clearly manifested in practice and amenable to a systematic review of classroom trials**. The field is yet to be systematically mapped and it was unclear at the point of protocol development what would be feasible within the resource envelope of the review. How tightly around the core focus of the review and what scope there is to include broader conceptions of cognitive science informed strategies for acquiring and retaining knowledge into the systematic strand of the review would depend on the—at that point unknown—weight of evidence from the more narrowly-defined core focus areas.
- Finally, the **definitions of cognitive science and the boundaries between related concepts such as metacognition vary between reviews**. As discussed above, the focus and scope of the EEF reviews on metacognition and self-regulation, social and emotional learning, and feedback (ongoing) form important considerations for the boundaries of the present review. One challenge in defining and managing boundary concepts will be to ensure duplication of effort with these reviews is avoided while highlighting essential areas of overlap and linkage.

These issues inform the methodology for this review, as detailed in Appendices 1 to 3. Navigation of these boundary issues involved the clear application of the review specification set out in the protocol and in particular:

- the use and application of clear inclusion and exclusion criteria;
- the application of a clear and specified search strategy;
- the use of coding frameworks to identify points of contact between the included studies, wider cognitive science, and educational moderating and contextual factors; and
- the exploration of issues that fall outside of the core systematic review criteria but have value in illustrating and interpreting the results (for example, variations in classroom practices that have not been trialled, or linkages between wider cognitive science concepts and those associated with specific strategies for acquiring and retaining knowledge); the 'evidence-informed discussion and questions' sections that follow the review of evidence in each area include this more exploratory aspect of the review.

Overview of review strands and their reporting

Review strands

The design of this review is underpinned by an appreciation that in translating cutting-edge research for application in education we are looking for a golden thread that runs through (a) the basic science, (b) the plausibility of its applications, (c) evidence of efficacy (of it working in controlled or ideal conditions), (d) evidence of effectiveness (of it working in normal conditions), and (e) the evidence of differential effectiveness by context or student groups.

Recognising this, our systematic review comprised a core strand and two further sub-strands:

- **core strand: systematic review**—a systematic review of published literature reporting cognitive science informed interventions inside classrooms; this process included assessing the robustness of the evidence (via eligibility criteria and, for selected studies, a 'risk of bias' analysis), the ecological validity of interventions, and their fidelity and relevance to (a) the cited underpinning cognitive science research and (b) the broader state-of-the-art in cognitive science (as per the basic cognitive science review);
- **sub-strand 1: basic cognitive science review**—a review of literature in contemporary cognitive science and practice-facing guidance literature that describes or explains the underpinning basic science research that is mobilised in interventions; and
- **sub-strand 2: practice review**—review of practice documents and school data collection to identify and illustrate the applications of cognitive science in the classroom and explore professional perspectives on the value, challenges, and application of cognitive science.

Objectives

Objectives: systematic review of cognitive science interventions in the classroom

Core strand: systematic review' – Systematic review of published literature reporting cognitive science informed interventions inside classrooms.

This core strand of work responds to the EEF research questions:

1. What is the impact within the classroom on pupil outcomes of approaches rooted in, or inspired by, cognitive science and have strong evidential underpinnings from cognitive science regarding memory and learning?
2. What are the key features of classroom approaches based on cognitive science that successfully improve pupil outcomes and teachers' and learners' contributions to them?
3. Do approaches rooted in, or inspired by, cognitive science have differential effects on outcomes for significant groups of pupils (for example, younger pupils or pupils eligible for free school meals) or in certain subjects? If so, what are the key features of those successful approaches?
4. What does this review tell us about how the five core strategies relate to the underlying cognitive science research, and each other?

We define cognitive science interventions through our conceptual framework, identifying the (a) classroom teaching and learning theories and (b) classroom teaching and learning practices under review. These are all informed by, or based on, cognitive science. We recognise that the nature and strength of the connection between cognitive science, theory, and practice vary.

We define 'effectiveness' as a difference in attainment between pupils receiving a cognitive science intervention and either (a) a business-as-usual condition or (b) a specified alternative condition that is causally attributable to the treatment conditions and is a suitable comparison condition for the main treatment condition given the strategy's theoretical and practical features.[3] Some alternative conditions are strategy-specific; retrieval practice, for example, is often compared to an equivalent restudy condition. The control conditions are identified in data extraction and accounted for in the

[3] That is, it compares two conditions that (a) put the theory supporting a cognitive science strategy to the test and (b) that teachers would consider to be reasonable alternative choices in normal classroom conditions.

analysis and reporting. We also will analyse effectiveness in terms of checking that no harms are being done (that the interventions do not lead to worse outcomes).

The security of causal attribution is judged using our quality assessment criteria, specified in the methods appendix (Appendix 1) and referred to in the substantive results sections.

We also assess the ecological validity of interventions reviewed in the core strand and their fidelity to (a) the cited underpinning cognitive science research and (b) the broader state-of-the-art in cognitive science (as per the basic cognitive science review). Again, criteria for this are specified below.

Finally, we note that the database produced as part of the core systematic review strand is designed in line with the coding frameworks and criteria of the EEF Evidence Database work, supporting the Teaching and Learning Toolkit. An additional objective for this review has been to support the database project and conduct this review following its methodology and aims.

Objectives: basic cognitive science review

Sub-strand 1: basic cognitive science review' – review of literature in contemporary cognitive science and practice-facing guidance literature that describes or explains the underpinning basic science research that is mobilised in interventions.

This sub-strand adds to the robustness of the review and deepens understanding, in education, of the underpinning cognitive science. This strand asks:

1. What is the state-of-the-art in cognitive science regarding memory and learning?
2. What is the current state of evidence about mechanisms for memory and learning and the effects of these mechanisms on learners and how confident is the field about this?
3. What are the links between the best evidence about cognitive science and about teaching and learning and what translation of this evidence for education, if any, has been tested or recommend by the field?
4. What does this review tell us about how the five core strategies relate to the basic science and to each other?

Examining the underpinning cognitive science and what educational implications have been recommended by the field allows this review to describe and assess how, and the extent to which, classroom interventions located in the systematic review strand are informed by specific areas of cognitive science. The definitions and discussions of the cognitive science areas and strategies in the main evidence review are informed by this strand of the review.

Objectives: practice review

Sub-strand 2: practice review' – Review of policy and practice documents, and school data collection to identify and illustrate the applications of cognitive science in the classroom.

The overall objectives of this practice review are to understand:

1. What applications of cognitive science in the classroom are currently prominent in policy, guidance, and practice? What do practitioners in England identify and recognise as common approaches based on cognitive science?
2. What form(s) do applications of cognitive science take when manifested in practice? How do cognitive science applications differ for different contexts, subjects, and groups of students?

We include full details of methods for the practice review in Appendix 13. In overview, it comprised a literature review and primary data collection.

Literature review: Alongside the main review, we reviewed literature to identify applications of cognitive science in the classroom from policy and practice documents (for example, reports, frameworks, guidance, and popular-scientific texts). The bibliography for this is provided in Appendix 2. The final bibliography, developed and accessed throughout the review, is provided in Appendix 13.

Primary data collection: We used interviews and a questionnaire to survey practitioners in England. Questions were developed as part of the practice review based on the questions above and refined following mid-point analysis from the core systematic review. Our survey was distributed via teacher and school organisations and social media.

Report structure

This report is structured as follows.

Part A: Review methods and background information

- In section **A1**, above, we provide **background information** for the review and introduce its **aims**.
- This is followed (above) by details of our **focus and problem statement**, the **review strands**, and their **objectives** (**A2**).
- In the next section (below, **A3**) we provide details of our **systematic review search strategy and results** for the systematic review. This provides an overview of the database we use for the main findings.

Part B: Evidence review

- The evidence review begins with an **introductory section** that describes the structure and content of each of the review sections. It summarises the approach to analysis and provides an overview of how the reporting is organised.
- This is then followed by **evidence reviews of eight areas of cognitive science** (sections B1–B8):
    - **B1.** Spaced learning
    - **B2.** Interleaving
    - **B3.** Retrieval practice
    - **B4.** Managing cognitive load
    - **B5.** Working with schemas
    - **B6.** Cognitive theory of multimedia learning
    - **B7.** Embodied learning and physical factors
    - **B8.** Mixed strategy programmes
- Within these sections, a total of **14 strategies are evaluated** using systematic review methods. A summary for each strategy and for each overall area is provided.
- The second part of each of the evidence review sections (B1–B8) is an **evidence-informed discussion and questions** section. This is an extensive non-systematic review and discussion of (a) wider evidence not assessed under the main strategy reviews,(b) scientific and practice-facing guidance literature pertaining to the theoretical principles and application of cognitive

science in the classroom, and (c) perspectives from practitioners from primary and secondary data collection and analysis from the practice review.

- The final section of Part B (**B9**) is a summary of **practitioner perspectives on cognitive science** from the practice review data (see above for details). All strategy-specific points are included in the evidence-informed discussion and question sections of the relevant evidence reviews (B1–B8). Therefore, section B9 is restricted to reporting general themes and perspectives from the practice review, drawing on a non-systematic literature review and the interview and questionnaire data.

Part C: Conclusions

- The first section in Part C is a **summary of results by area** (**C1**). This section collates and repeats the individual summaries by strategy and by area in the main evidence review sections (B1–B8).
- Section **C2** then presents the **overarching findings** of the review and their **implications** for research, policy, and practice. This includes consideration of whether the results are in agreement with a previous review of cognitive science in the classroom.
- The **limitations** of the review are detailed in section **C3**.
- Section **C4** provides **information about the review**, including the review team and advisory group.
- Finally, **C5** provides **references** for the main report and discussion sections in the evidence reviews in Part B. Note that references for the studies in the main review database used in the systematic reviews in sections B1–B8 are included in dedicated appendices (5 to 12).

A3. Systematic review literature search

Introduction

This section describes the search strategy and results for the main systematic review strand of this review. The systematic review located and reviewed published literature reporting cognitive science informed interventions inside classrooms.

Search strategy and results

Our search strings were based on the concepts identified on our conceptual map (Appendix 2) and selected following scoping work described above. We developed our initial search terms through preliminary database searches to assess search term sensitivity and precision[4]. We also considered feedback from advisory group members (see Appendix 1) about prioritising and defining cognitive science concepts.

We designed our searches to include terms related to a) methodology, b) education (outcomes and classroom specific), and c) terms and synonyms related to the specific cognitive science area (including a general cognitive science search). Table A3.1 and A3.2 provide the search term fragments that were combined to create our search strings. These search terms were entered into each search database with the minimum of adaptation needed to use the search syntax and functionality and ensure comparability across databases. We applied these across ten databases (including 20 collections), as set out in detail in Appendix 1.

Table A3.1: General search terms (all searches)

Search term group	Search string (fragment)	Search location[1]
Group 1: methodology	intervention OR trial OR evaluat* OR experiment* OR quasi-experiment* OR pilot OR test*	Title, abstract, or key words
Group 2: education outcomes	AND learning OR attainment OR achievement OR 'test scores' OR outcomes OR exam* OR impact OR effect OR performance	
Group 3: classroom setting	AND classroom OR teach* OR school OR 'further education' OR nursery OR 'early years' OR kindergarten OR pre-primary OR lesson	
Group 4: focus concept	AND, one of the general or concept-specific search term fragments in A3.2, below.	

[1] Subject to search database functionality.

The general search terms above will be combined with one of the search strings related to cognitive science in general and specific cognitive science concepts, below.

[4] https://handbook-5-1.cochrane.org/chapter_6/6_4_4_sensitivity_versus_precision.htm

Table A3.2: Cognitive science concept-specific search terms—core concepts

Cognitive science concept	Search string (fragment—to be combined with the general search terms, above)	Search location[1]
Cognitive science general	cog* OR brain* OR neuro* OR 'learning science'	Title, abstract, or key words
Spaced practice	spac* OR distributed	
Interleaving	interleav* OR interweav*	
Retrieval practice	retriev* OR 'testing effect'	
Dual coding	dual	
Strategies to manage cognitive load	'working memory' OR 'short-term memory' OR (load AND (Cognitive OR intrinsic OR extraneous OR germane))	

[1] Subject to search database functionality.

Records located from searches

Our initial database contained 41,125 records. This was the figure after the removal of duplicates (n = 7,615) and including the additional records identified from the references of other reviews (n = 377). This vast number reflects the number of searches for the extensive range of concepts and strategies within the scope of the research. Moreover, many of the search terms were deliberately general, prioritising identifying as many relevant studies as possible over precision and efficiency. As a result, a very large number of studies were excluded through a title and abstract screening (see below).

Inclusion and exclusion criteria for the review

Our approach to searching and screening was iterative, with two overall groups of eligibility criteria to be applied. The **initial eligibility criteria** for studies to include in the core systematic review, as per the review protocol, are given in Table A3.3.

Table A3.3: Staged application of initial eligibility criteria.

	Round 1 Screen titles and abstracts	Round 2 Screen full reports
1. **Population**: Children and young people between 3 and 18 years of age in classroom settings.	✓	✓
2. **Interventions/practices of interest**:		
i. Evaluation of a classroom trial and/or intervention.	✓	✓
ii. Uses approaches derived from cognitive science relating to the acquisition and retention of knowledge.	(✓)[1]	✓
3. **Study design and outcomes**:		
i. Initially include all studies reporting empirical evidence of any type or quality about pupil impact, including reviews, which we 'mined' for underpinning studies.	(✓)[1]	✓
ii. Studies which have any form/quality of counterfactual.	(✓)[1]	✓[2]

iii. Flag (but exclude) reviews and meta-analyses for reference mining and to inform the basic science or practice review strands.	✓	✓
iv. Flag (but exclude) pieces that are of relevance to the basic science or practice review strands.	✓	✓
4. **Language**: Include pieces written in English and peer reviewed (for journal articles).	✓[3]	✓[3]
5. **Bodies of Literature**:		
i. Include all peer reviewed journal articles and reports based on research commissioned by policymakers, charitable or other non-commercial organisations.	✓[3]	✓[3]
ii. Exclude conference proceedings, working papers and master's and doctoral dissertations/theses that were published before January 2017.	✓[3]	✓[3]

[1] Assessing this item was to some extent possible from title and abstract screening, with definite 'no's' being removed.

[2] As discussed below, a decision was made about level of stringency for the study design and quality criteria following an initial literature mapping after round two (see below).

[3] These final criteria were mostly applied during database searching but remained as eligibility criteria during screening for any records for which initial information was missing or erroneous.

Following an appraisal of the coverage and quality of evidence (see below) across the various cognitive science areas, we tightened the criteria for study design (3, above) to include only experimental and quasi-experimental studies (see Appendix 1 for further details of how we defined this).

After this initial screen, we mapped and categorised the database (see further details below). We then applied a **second eligibility assessment** using the following criteria. The final eligibility assessment tool is provided at the end of Appendix 1. This was informed by discussion with the project advisory group and was designed to identify the evidence with the greatest potential relevance and quality for the review. We assessed each study as being 'high', 'medium', or 'low' eligibility in four areas:

1. **Relevance and definition for focus cognitive science practices**. This assessed (a) the study's relevance to our cognitive science strategy definitions and focus questions and (b) the strength and clarity of the test of the strategy and/or principle. For this we looked for a clear and relevant counterfactual and controlled conditions. Relevant counterfactuals are strategy-specific: each cognitive science concept implies alternative strategies that are not aligned with the principle in question, for example, massed versus spaced practice, restudy or re-presentation versus retrieval practice, and so on (see definitions in each of the evidence review sections). For purposes of transparency, several studies not meeting this criteria are detailed and indicated in the overview of studies for each strategy, but not included in the results. The requirement to have controlled conditions extended the design criteria—3, above, concerning the need for experimental or quasi-experimental designs—to also require defined interventions or conditions that would test a cognitive science strategy or principle. There were studies, for example, where a cognitive science strategy or principle was an incidental or minor aspect of a study designed to examine another question. The need for this second stage of assessing relevance stemmed in large part from the challenges of operationalising the concept of 'cognitive science informed' intervention. This concept did not lend itself to pre-specification and needed to be assessed against the actual data.

2. **Ecological validity**. For this, we developed a specific tool that assessed (a) system relevance, (b) research context (that is, where and by whom the research was conducted), (c) cohort size (in terms of the number of pupils, teachers, and schools involved), and (d) learning outcomes (in terms of whether the learning outcome was a typical curriculum objective as opposed to learning in a recreational game or non-curricular psychometric assessment. Each of these criteria had low/medium/high levels with a level descriptor. A best-fit overall level was given for ecological validity from these sub-criteria.
3. **Added value to evidence-base**. This was an item introduced after advisory group discussion. It was designed to add breadth to the review by flagging studies that offered evidence for under-represented factors in the literature. This was a category-by-category judgement based on an assessment of coverage in the group of studies (after mapping) in terms of pupil age, subject area, disadvantage status, type or aim of learning activity or instructional relation, and training or resource requirement. In practice, this criterion was largely used for information purposes to identify studies that added breadth to the database. It had the practical effect of retaining a small number of borderline studies in terms of relevance and ecological validity that otherwise would be excluded. These studies were in under-represented contexts such as early years and arts and humanities subjects. The inclusion decision was made prior to the extraction and analysis of results.
4. **Overall rating**. This was the best fit judgement across all categories and was based almost entirely on the first two criteria. In short, we assessed whether a study was 'testing a cognitive science informed strategy in realistic classroom conditions'. The assessments of the individual criteria provided more granular information to inform and record this judgement.

Screening process

Screening

After screening the database of 41,125 records on title and abstract, we were left with n = 2,193 studies. After another round of screening, using the same eligibility criteria applied to the full text, we were left with 700 studies. Many of the original records were excluded because they were not a study of pupils aged 3 to 18 in a classroom setting (many abstracts did not include this information and, on inspection of the paper, we found, for example, that the study was conducted with undergraduates). There were also many examples of studies that could have been relevant to our cognitive science concepts from the abstract but the full paper did not bear this potential out. Again, we sought as comprehensive a database as possible and erred on the side of caution, including papers with potential eligibility rather than demonstrable eligibility until the full text could be reviewed. As per the protocol, we double-coded 20% of records at each stage of screening (including the eligibility assessment, below). We double-coded the first 20% of all records. Discussion and reconciliation of judgements following this was designed to resolve disagreements for specific records and improve inter-rater reliability on all subsequent items. Two researchers (RL and TP) coded these records and another (CJ) adjudicated any disagreements. Further details on this are provided in Appendix 1.

Categorisation

At this point, we categorised papers by cognitive science area. Initially we used the five core strategies identified above. We then expanded the categorisation to a more granular list containing closer to 20 sub-areas. We then regrouped the papers back to eight categories (as per the evidence review sections in Part B, below). The original strategy areas were mostly intact, although we generalised dual coding

to the Cognitive Theory of Multimedia Learning (section B6) and added a 'mixed strategy' group, an 'embodied learning and physical activity' group, and a 'working with schemas' group.

Eligibility assessment

Following categorisation, we conducted a full eligibility assessment of the 700 papers. This applied the second set of eligibility criteria described above, including the use of our newly-developed ecological validity tool. The primary aim of this was to organise the studies into those with high, medium, and low priority for the review (in terms of their potential to answer our review questions). Given the size of the database at this point still being unfeasible, we tightened the definitions of 'cognitive science' and 'ecological validity' during this process. The latter was tightened via the use of the ecological validity screening tool (see Appendix 1), which went beyond the population, setting, and outcome criteria from the initial screen. The cognitive science relevance was also tightened from the initial screen. As noted above, 'cognitive science informed intervention' was a challenging concept to operationalise. In our first round of screening, we erred on the side of caution, retaining studies with more tenuous links to cognitive science or vaguer operationalisation and testing of cognitive science strategies. Having initially looser interpretation of the criteria and then tightening allowed us to build up familiarity with the evidence-base and the borderline-eligible studies, enabling us to be more confident of consistency when applying the tighter criteria. We described the process and reasons for the need for iterative application of criteria in the original protocol. In effect, the second round of eligibility assessment organised studies into four groups, 'high', 'medium', and 'low' priority, and excluded a number as a result of tighter relevance criteria and the use of a more precise ecological validity assessment tool. This stage resulted in 201 more exclusions. We also identified 20 duplicates (these were not identified as duplicates by the software but by hand). The vast majority of these were studies sourced from other reviews that were added after initial duplicate removal. **The final database contained 499 records**, the vast majority reporting substantive and distinct studies.

Selection of studies for analysis

The 499 studies in the final database were prioritised for inclusion in, and treatment by, the systematic review based on their ratings against the above criteria. Of these, 43 had a high priority rating, marking them to have high potential value as a test of cognitive science in the classroom, 252 were rated as medium priority, and 204 were rated as low priority. The low priority studies were not included in the analysis so were effectively excluded. The high priority studies were analysed in-depth. Table A3.4 provides the number of studies in the review database for each rating (high, medium, and low) for each of the **second eligibility assessment** criteria (as above).

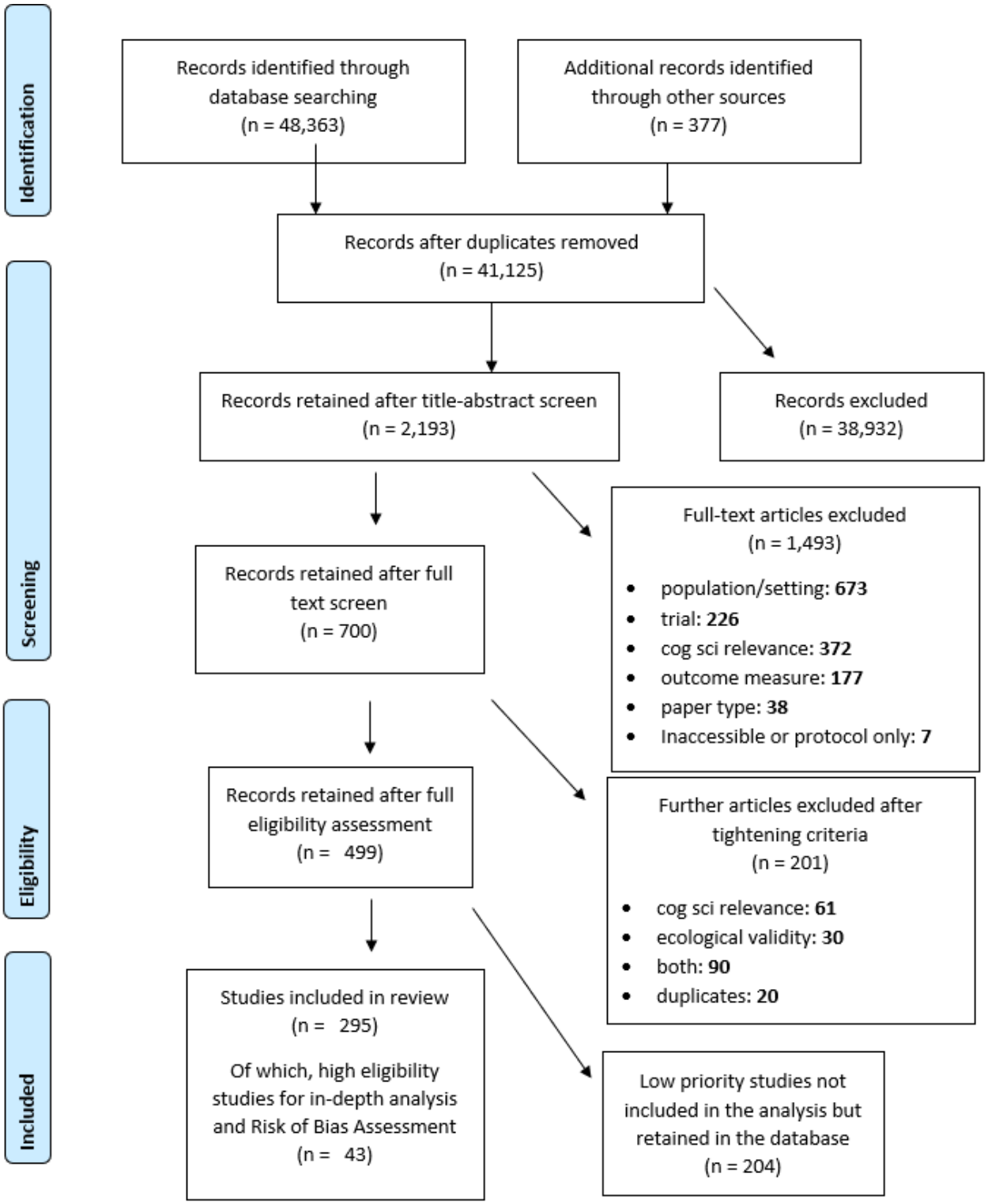
Table A3.4: Overview of all database records and results of the priority assessment

Priority level	Overall rating	Ecological validity	Relevance and definition for focus CS practices	Added value to evidence-base
High	43	45	141	97
Medium	252	227	268	322
Low	204	227	90	80

Further details of the treatment and analysis of high and medium priority studies are included in the main evidence review sections along with a breakdown of the priority assessment results for each area. We provide a PRISMA diagram giving an overview of the screening and priority assessment

process in Figure A3.1 and this is reproduced in Appendix 3. Please see Appendices 1–4 for further details of the review methods.

Figure A3.1: PRISMA flow diagram overview of screening and eligibility assessment



Part B: Evidence review

Evidence review introduction

Overview of the evidence reviews

Organisation of reporting

Part B of this report presents all of the evidence we have reviewed. This includes:

- a systematic review of trials of cognitive science in the classroom;
- a non-systematic review of wider evidence, the basic science, and practice-facing guidance; and
- primary evidence from practitioner interviews and a questionnaire.

We have adopted a standardised, systematic structure to present this evidence and ensure clarity about the provenance of the evidence and our approaches to analysis. This structure is designed to allow a rigorous, transparent test of the overall cognitive science strategies that have been tested in the classroom while also recognising and allowing exploration of theoretical and practical evidence. The wider theoretical and practical evidence is—by its nature and due to the limitations in the present evidence-base—challenging to review systematically; this evidence, however, provides an important part of the account of the present state of knowledge and the considerations and uncertainties faced by practitioners and researchers as they continue to advance our knowledge of cognitive science in the classroom.

Part B comprises **evidence reviews of eight areas of cognitive science** (sections B1–B8) followed by a single section summarising **practitioner perspectives on cognitive science** from the practice review literature and data (B9). Note that the *area-specific* practitioner perspectives (for example, on retrieval practice) are included in the discussion sections of the main reviews (B1–B8); the dedicated section B9 focuses on general and cross-cutting perspectives.

The cognitive science areas reviewed are:

- **B1.** Spaced learning
- **B2.** Interleaving
- **B3.** Retrieval practice
- **B4.** Managing cognitive load
- **B5.** Working with schemas
- **B6.** Cognitive theory of multimedia learning
- **B7.** Embodied learning and physical factors
- **B8.** Mixed strategy programmes

In the first half of each of these sections, a total of **fourteen strategies are evaluated** using systematic review methods. A summary for each strategy and for each overall area is provided.

The second part of each of the evidence review sections (B1–B8) is an **evidence-informed discussion and questions** section. This is an extensive, non-systematic review and discussion of:

a) wider evidence not assessed under the main strategy reviews;
b) scientific and practice-facing guidance literature pertaining to the theoretical principles and application of cognitive science in the classroom; and

c) perspectives from practitioners from primary and secondary data collection and analysis from the practice review.

On overview of what the reader can expect in each evidence review section is provided in Box B.1.

Box B.1: Reporting of findings in evidence review areas, by section

Overview of area

In Part B, we review 14 cognitive science informed strategies organised into eight review areas (for example, spaced practice, managing cognitive load, and mixed strategy programmes). Each of the eight areas begins with a general discussion and overview, including definitions of the strategies and principles being assessed in the section.

***These overview sections have drawn on literature from across the review** and, in particular, sources from the practice review and basic science review when introducing and defining each strategy and concept.*

Main findings

***The evidence reviews for each of the 14 strategies in the eight main review areas are based solely on analysis of studies from the core systematic review strand* only**. *All studies have met our eligibility criteria in full. Studies have been analysed and reported using systematic review methods. This includes the 'Summary of Findings' for each strategy as well as the 'Overall Evidence Summary' for each area of the review.*

Evidence-informed discussion and questions

The results from the systematic review in each of the eight areas are followed by a section described as 'Evidence-Informed Discussion and Questions'. In this section we (a) discuss the main results and (b) pose questions that have value for practice and future research for which we cannot provide a robust answer given the evidence we have.

***This section draws on all review strands**, connecting (a) the main results from the systematic review of applied cognitive science, (b) the principles and evidence from the basic science review, and (c) the empirical evidence and literature review from the practice review. In this section we bring together theory and evidence from basic science, applied science, and practice.*

Analysis of evidence

Priority—identification of studies for analysis and in-depth analysis

The main findings are based on analysis of the subset of studies identified as high and medium priority (see 'Selection of studies for analysis', page 24).

- High priority studies were deemed to have high *potential* as a basis for judgements about strategy effectiveness. Key information for these studies was extracted into data tables (for which an overview is provided in each strategy results section and the corresponding appendix). Additionally, high priority studies were reviewed in-depth and a full 'risk of bias' assessment was conducted, summarised in each strategy results section. Where risk of bias was judged to be low, high priority studies have been given additional weight when assessing the evidence for each strategy.
- Medium priority studies were included in the analysis for each strategy. Key information for these studies was extracted into the data tables. These studies were assessed collectively in conjunction with the high priority studies for the strategy.
- Low priority studies were not included in any analysis. While these studies had met broad eligibility criteria in the first round of screening, their 'priority' in terms of relevance, ecological validity, and potential to add value to the evidence-base was low. These studies are characterised as being small, conducted in circumstances that were somewhat atypical of everyday classroom conditions and learning, and with only tenuous links between the study and our focus cognitive science strategies.

Through this prioritisation, we aimed to strike a balance between a focus on in-depth analysis of evidence with the greatest potential to evaluate cognitive science strategies (high priority) and a broader analysis of a wider evidence-base, which provides more indirect (that is, lower relevance and ecological validity) evidence. In total, there were only 43 high priority studies across all eight areas and strategies (Table A3.4). If we had restricted our analysis to these, the weight of evidence for most of the strategies would have been too limited to reach a judgement on. By also including the medium priority studies (252), we were able to draw conclusions for most strategies. It was not feasible to conduct a full risk of bias analysis for the medium priority studies; also, the lower relevance and ecological validity of these studies required more judgement to assess. See below for further details of the systematic process we followed during analysis.

Strategy grouping and selection for analysis

Prior to analysis, all high and medium studies in the area were grouped by strategy. Where there was a sufficient weight of evidence (that is, a sufficient number of high studies or groups of medium studies) we evaluated the evidence for the strategy.

Where the weight of evidence for a strategy was *not* sufficient to evaluate the evidence, or the studies had a too-specific or theoretical focus, they have been incorporated into the Discussion and Questions section where we pose and discuss questions for theory and practice in this area (see below for further details). For example, there were studies that were designed to examine the conditions in which strategies worked that compared variants of the treatment (for example, different lengths of spacing) or strategies or variants of our focus strategies (for example, the use of 'faded' worked examples). Many of these studies were concerned with *how* strategies worked rather than *if* they worked or were too few in number to constitute a viable strategy group for analysis.

Main analysis

Within results sections for each cognitive science concept area, the analysis is both conducted and reported via a combination of the following.

- A **characterisation of the evidence in a strategy group** in the following areas:
  a. **Pupil age and characteristics**—focusing on the age of pupils represented by the studies and noting any particular characteristics of pupils or settings that may influence the results. Few

examples of the latter were identified and reported as most studies were from mainstream, mixed educational settings.

b. **Location**—reporting the country location of the study. Most of the literature included in this review comes from high-income countries. It is as argued by Abdazi (2014) that it is important not to assume that strategies work equally well or are equally suitable in low-income countries. Cognitive science is not equally embraced everywhere (Aronson, 2020). We report the location of studies for interested readers or those working in particular contexts where this information is relevant, however, we note that the country context of the study was not considered in the analysis.

c. **Learning areas**—here we report the subject areas covered by the studies in the group. Where information on specific curricular topics or learning objectives is provided and relevant, we also report this.

d. **Outcome measures**—focusing on the outcome measure assessments used in the research. Our primary reason for providing this information is to distinguish between high-validity, standardised educational assessments in the relevant subject area, and narrower or researcher-designed assessments that may inflate effect estimates or lower the confidence that the results will transfer to other measures.

e. **Design and delivery**—focused on ecological validity and whether a study was delivered by teachers in everyday classroom conditions. In many cases, studies were conducted by researchers or made use of scripted lessons or learning software. We also note, where relevant and the information was provided, whether teachers received professional development or resources to support intervention delivery.

- An **overview table of all studies** in the analysis for the strategy. The overview tables are structured to present the high priority studies first; this is followed by the remaining medium priority studies organised by size (with studies with more than 500 pupils being deemed large, studies with 101 to 500 pupils medium, and those with less than 100 pupils small). We provide effect size estimates where these were reported in the original papers or we were able to calculate them. We also provide a study-level overview summary of the findings as 'positive', 'neutral', or 'negative' in terms of whether the results suggested a positive effect of the cognitive science strategy in question relative to a business-as-usual control condition or strategy-specific alternative. We urge readers not to 'vote count' these study-level summaries.

Our main analysis is based on results from the following two tools:

- An **adaptation of the GRADE approach** to assessing the certainty in, and limitations of, the evidence for each strategy group. As we describe in more detail in the limitations section (C3) and the Methods appendix (Appendix 1), the GRADE approach is geared towards quantitative summary and assessment of evidence quality. The weaknesses of the evidence-base and constraints of review resources for quantitative coding of information necessitated a more qualitative, narrative summary approach, while still reporting on the key features required within a GRADE analysis, which we interpret as follows:

  a. **Risk of bias**—summarising risks identified for high priority studies. We also report study designs in the group.

  b. **Inconsistency**—when there is significant and unexplained variability in results from different trials.

  c. **Indirectness**—where narrowness in the population or intervention, or lower adherence to a cognitive science strategy definition, causes the evidence to be a more indirect or weaker test of the focus strategy.

d. **Imprecision**—when wide confidence intervals, small studies, or low-validity measures reduce confidence in the impact estimates.
e. **Publication bias**—where we looked (without a formal quantitative assessment) for evidence of publication bias (for example, with larger studies providing lower estimates than smaller studies).
f. **Other considerations**—where we comment on particularly strong studies with large effects, studies that examine dose-response, or studies that present evidence supporting the underlying theory (for example, also measuring cognitive load to show that the intervention has the intended impact on working memory as well as outcomes).
g. **Overall confidence**—where we report the certainty of the evidence on one of four levels, described by the following standard descriptors:
    i. **high**—we are very confident that the effect of the study reflects the actual effect;
    ii. **moderate**—we are quite confident that the effect in the study is close to the true effect, but it is also possible it is substantially different;
    iii. **low**—the true effect may differ significantly from the estimate; and
    iv. **very low**—the true effect is likely to be substantially different from the estimated effect.

We provide an overall confidence level and an explanation of why the evidence group was down- or up-graded to the final level. We also comment on other considerations where relevant to the strategy area and evidence.

- A **structured narrative summary** of the evidence. This drew on the overview table and the summary table of the GRADE assessment to summarise the results in the following five areas:
    a. **main finding**—summarises results from approaches one to three above;
    b. **estimated impact**—summarises strategy effect size estimates, noting the size, priority, and any risks of bias for the studies;
    c. **confidence in impact estimate**—summarises the GRADE assessment;
    d. **heterogeneity—**comments on variation identified in the evidence characterisation and raise in the GRADE analysis that are relevant to the interpretation of the findings; and
    e. **other points**—notes any specific points relevant to the strategy, analysis, or evidence.

The combination of these three approaches to analysis and the transparent reporting of this represents a significant advance on narrative synthesis approaches typical of systematic reviews in education research. Nonetheless, there is inevitably a substantial degree of subjectivity (that is, use of expertise and judgement) involved in this approach. Judgement is required for all systematic reviews across all fields (Gough, Oliver and Thomas, 2017); cognitive science in the classroom and the complex, nascent and often disparate nature of the applied evidence made this particularly so. Challenges in the analysis stemmed from the challenge of systematically navigating the appreciable variation evident in relation to:

- the strategies in a given group;
- the populations and contexts in which they were tested;
- the choice of outcome measure;
- learning objective, topic, and subject area;
- comparisons and research questions tested;
- study size and quality;
- level of ecological validity (and therefore adherence to our eligibility criteria requiring tests in realistic classroom conditions);
- quality of the reporting; and

- relevance and adherence of studies to the definitions (or our definitions) of the cognitive science areas and strategies.

As discussed further in our Methods appendix (Appendix 1) and the limitations section (C3), the challenges and limitations of attempting a fully quantitative approach to coding and analysis led us to a mixed-methods approach drawing on both quantitative and qualitative information. We believe that the present analysis strikes a good balance between several constraints and provides the most informative, valid, and productive contribution to this developing field at this time. Moreover, the analysis was designed to be as transparent as possible to enable any biases or misjudgements to be apparent to readers, and readers to be in a position to draw different conclusions.

Evidence-informed discussion and questions

As described above, where the weight of evidence for a strategy was *not* sufficient to evaluate the evidence, or the studies had a too-specific or theoretical focus, they have been incorporated into the Discussion and Questions section where we pose and discuss questions for theory and practice. In these sections, we present a non-systematic review of theory and evidence, drawing on:

- *all* high and medium studies (including from the main systematic review and studies not grouped under the assessed strategies);
- selected sources from the basic science literature to explain and examine the underpinning neuroscientific and psychological bases of the focus strategies;
- wider sources from the bibliography of the practice review, including practice-facing guidance literature; and
- data from questionnaires and surveys from the practice review.

For the majority of sections, we have grouped this discussion into three sections:

- evidence about the theoretical principles and mechanisms that are thought to underpin effective use of the strategies;
- variations in the cognitive science practices in focus, the teaching and learning contexts in which they are used, and the pupil groups and aims they are relevant for; and
- key implementation factors, barriers, and facilitators.

It is important to note that, by design, all discussion in this section is of strategies for which there was *not* a sufficient weight of evidence to evaluate effectiveness. This was because of heterogeneity within studies prevented grouping by strategy or that the nascent nature of the evidence-base did not enable evaluation. However, these studies provide helpful detail, texture, and context for the main findings. At present, theory and perspectives on effectiveness by far outstrip the evidence-base from applied classroom trials. Reflecting this weakness in the evidence-base, this discussion section positions the theory as working hypotheses for research and practice. It poses and discusses evidence-informed questions for practitioners who wish to implement the strategies and for researchers as they work to develop and test the theory.

B1. Spaced learning

Overview of area

Definitions

'Spaced learning' applies the principle that material is more easily learnt when separated by an inter-study interval (ISI). ISIs can be very brief (seconds or minutes) or very long (weeks or months). Spaced learning is also referred to as 'spaced practice', 'distributed practice', 'distributed learning', and the 'spacing effect'. A number of scientific theories have been proposed to explain the benefit of spacing for long-term retention, though they often lack substantiation in applied contexts. One school of thought proposes that ISIs may facilitate the consolidation of memories—the formation of new knowledge gleaned from the learning of new subject material (Smolen, Zhang, and Byrne, 2016). Spaced practice is often contrasted with massed (or clustered) practice, whereby material is practiced in a single session or close succession. Spacing spreads out study activities over time (Dunlosky and Rawson, 2015) and can be implemented in several ways. Material can be spaced within a single lesson, for example, by revisiting a new concept three times in a lesson with ten-minute spaces in between. Alternatively, spacing can occur between lessons—a topic may be revisited three times across one week or once a week for several weeks. In the literature, there were more examples of spacing across days and weeks; we therefore refer to these as 'standard spacing' or 'spacing over lessons'. There were fewer examples of spacing within a single lesson; we refer to these as 'short' or 'within-lesson' spacing.

Spacing can be applied to many aspects of teaching and learning, including the spacing of instruction or delivery (for example, information provided on a particular topic), practice (such as completing worksheets), or assessment (for example, the frequency of quizzes or formative tests). Spaced learning is one of several cognitive science informed strategies labelled as a 'desirable difficulty'; learning may be more challenging on a short-term basis, but long-term retention is thought to be enhanced as a result (Greving and Richter, 2019). In spaced practice, spaces are usually filled with unrelated activities or the learning of unrelated topics.

Overview of the evidence-base

Table B1.1: Spaced practice—overview of study priority ratings

Priority level	Overall rating	Ecological validity	Relevance and definition for focus CS practices	Added value to evidence-base
High	5	7	34	8
Medium	22	15	9	32
Low	19	24	3	6

The review study database contained 46 studies in the spacing category. Of these, 27 were graded as having sufficient ecological validity, relevance, and value for inclusion within this analysis of the evidence (high and medium priority). Five studies scored highly across these criteria and were identified as *potentially* providing strong evidence in this area (high priority). Relevance and adherence to the definition was strong (as indicated by 34 of the 46 studies being rated high). Spaced practice is readily defined and identified, and locating studies testing this principle and grouping trials of homogenous practices was relatively successful.

In this area we have identified two strategies with sufficient evidence to examine the effectiveness of the strategy. These are defined more fully below. The two strategies are:

- **'standard' spaces, across lessons or days**—18 studies, of which three were graded as high priority and thereby identified for in-depth analysis; and
- **'short' spaces, within lessons**—two studies, both high priority, one a meta-analysis of six small-scale spaced learning trials in this area.

As detailed in the introduction to Part B, not all high and medium priority studies formed a sufficiently large group to enable the strategy to be evaluated in the systematic review. In this case, there were 27 high and medium priority studies, of which 20 have been grouped into the two strategies above. Ungrouped medium and high priority studies form part of the wider evidence discussed in the 'evidence-informed discussion and questions' section, as explained in the introduction to Part B. Wider evidence in this area looks at personalising spaced practice, combinations of massed and spaced practice, practice over a longer period of many months, moderation by learner analytical ability, and possible links between spacing and working memory depletion.

Main findings

Strategy 1: spacing across days or lessons ('standard' spacing)

Concise definition

'Standard' spacing is the practice of separating or distributing learning over more than one lesson, usually across multiple days and weeks.

Full definition and description

'Standard' spacing is the practice of separating or distributing the (re)presentation or (re)study of material over more than one lesson, usually across multiple days and weeks, in some cases over longer periods. Spacing is sometimes called 'distributed' learning or practice. The alternative to spacing is usually referred to as 'blocked' or 'massed' practice, where content is studied in a single learning session. Spacing is sometimes combined with retrieval practice, often known as spaced retrieval practice.

Selected examples

Examples of this strategy from our database:

- Denton et al. (2011) delivered a supplemental, small-group reading tutoring intervention to first-grade students (age 6 to 7) over (a) four sessions per week for 16 weeks, (b) four sessions per week for eight weeks, or (c) two sessions per week over 16 weeks). The first group received approximately 30 hours of the intervention and latter two groups about 15 hours each (and so could be compared to assess spaced practice).
- Bloom (1981) compared ten minutes on three successive days to 30 minutes on a single day of vocabulary practice in a high-school French language course.
- Goossens et al. (2016) compared six exercises a week to the same spread over two weeks for a primary school vocabulary learning study.

- Greving and Richter (2019) compared recall of a test for seventh-grade students (age 12 to 13) who re-read a text immediately compared to re-reading it one week later.
- Svihla et al. (2018) studied high school student's learning in a science enquiry unit. One group completed the unit in five consecutive class periods over two weeks; the other group completed one activity per week for five weeks.
- Küpper-Tetzel, Erdfelder and Dickhäuser (2014) compared learning in German-English vocabulary pairs from two sessions separated by zero, one, or ten days.

Evidence for this approach

There were 18 studies for 'standard' spacing across days or lessons. Of these, two were identified for in-depth study. Full details of all medium and high priority studies are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed for the standard spacing strategy are characterised as follows:

- **Pupil age and characteristics**. The age range of students was first grade (Year 2, age 6–7) to 11th grade (age 16–17). There was a good spread of ages across this range, although for the youngest children (first grade) there were only two small studies and no studies for children any younger than this.
- **Location**. The majority (seven) of the studies were conducted in the U.S., with five from Germany, three from the U.K., two from Netherlands, and one from Iran.
- **Learning areas**. There were five studies focused on improving reading and vocabulary, these were all for children of primary age; there were three studies of second language vocabulary, all for secondary-age pupils; there were four studies of secondary science, of which two were focused on biology, one on science in general, and another on inquiry in science; there were four studies focusing on maths that spanned the full age-range from primary school multiplication and vocabulary to secondary statistics and probability; one study focused on critical thinking for 10- to 14-year-olds; finally, one focused on high school volleyball skill.
- **Outcome measures**. Many of the interventions relied on researcher-developed measures and focus on learning content-specific vocabulary tests. Similarly, reading tests were focused on recall tests of letter-sound knowledge. The tests of critical thinking, maths, and science inquiry were also researcher-developed. The volleyball skill test was based on a standardised test from a national organisation. For the science tests, one used content from the national exam boards for GCSE biology (Feddern et al., 2018) and one (O'Hare et al., 2017) a national exam board past paper. In overview, tests of learning outcomes are—for all but a very small number of studies—limited to simple recall items on research-designed and content-aligned tests.
- **Design and delivery**. The vast majority of studies were based on lessons or practice sessions designed by researchers, sometimes using online resources or computer programmes to provide lesson plans, materials, or study guides. Many of these also had lessons delivered or supervised by researchers. Several studies were delivered by teachers, with training, guidance, or materials provided by researchers (Bloom et al., 1981; Denton et al., 2011; French et al., 1990; Svihla et al., 2018; O'Hare et al., 2017; Sobel et al., 2011). Overall, ecological validity was moderate, with studies focusing on testing spaced practice principles in real classrooms, but often neither with teachers nor as part of the regular curriculum.

High priority studies in this area

There were three studies in the standard spacing strategy category that were rated as high priority (see above for details). We conducted an in-depth analysis of these studies and have completed a full risk of bias assessment (summarised in Appendix 5).

Feddern et al., 2019. This study employed a randomised controlled trial (RCT) to test the effectiveness of biology revision software incorporating cognitive science principles on biology test scores. The trial involved 14-year-old pupils in a U.K. school (n = 829). There were three conditions: first, a 'business as usual' group who completed a 40-minute revision session using a physical guide (massed practice); second, an 'offline' spacing group who completed two 20-minute sessions using a PDF revision guide two weeks apart; third, a software condition using mixed cognitive science strategies and question personalisation based on performance. Students worked independently. The learning measure was a pen-and-paper biology test on the content, consisting of multiple choice, free recall, and short answer questions. Although several principles were included in the latter condition (that is, spacing, interleaving, retrieval, and visual cues), it is possible from the first two conditions to analyse the spacing effect compared to massed practice specifically. We return to the mixed strategy condition in our evidence review of mixed strategy interventions.

Key findings. 'Offline' spacing (mean score = 5.15) was found to be slightly, but not statistically, significantly more effective than massed practice (M = 4.08); the mixed strategy software condition (M = 8.39) produced significantly higher scores than the spacing and massed conditions. As the software condition included several other strategies and spacing, we base our spacing results on comparing only the first two conditions. The mixed strategy has been reported in this section solely for comparison purposes (see Mixed Strategy Evidence section for further details of the mixed strategy software condition). The risk of bias assessment for this study did not raise any concerns. However, we note that this was published in the Chartered College of Teaching Impact journal and was at a shorter length, with briefer and less formal reporting, than typical of a journal with a research audience. This also meant that information such as test score standard deviations by group were missing.

O'Hare et al., 2017.This second study rated as high priority in this area. It was an RCT pilot evaluation of the EEF SMART Spaces programme for 13- to 15-year-olds in England (n = 408) on GCSE science test performance. The groups were allocated at the class level into three experimental groups and two control groups. The conditions were:

- three spaced practice experimental groups:
    - version 1: ten-minute spaces within class (n = 110),
    - version 2: 24-hour delay, interleaved topics (n = 75), and
    - version 3: both (n = 91); and
- two control conditions:
    - a 'slides-only' control, where the materials were provided (n = 79); and
    - a 'business as usual' control (n = 53).

After training, teachers were supplied with curriculum-based lesson materials for three topics, one each for biology, chemistry, and physics. The trial was conducted over three consecutive days, with intra-lesson spacing and/or inter-lesson spacing manipulated. A GCSE past paper was used as the outcome measure, with secondary outcome measures examining pupil engagement. There were both short-answer and long-answer test items.

Table B1.2: Summary of results from O'Hare et al., 2017—total test scores

Variant	Control	Effect size (g)	95% CI
10-min	Slides only	0.03	(-0.14, 0.20)
	Business as usual	0.12	(-0.07, 0.30)
24-hr	Slides only	-0.09	(-0.25, 0.07)
	Business as usual	-0.02	(-0.20, 0.14)
Both	Slides only	0.11	(-0.05, 0.28)
	Business as usual	0.19	(0.01, 0.36)

Key findings. The study found that *combining* 24-hour and ten-minute spacing was most effective (particularly on long-answer questions) compared to BAU, although the effect was small (*g* = 0.19, 95 % CI: 0.01, 0.36). This effect was the largest; all others were not statistically significant from zero. A pupil questionnaire included a scale-based measure of pupil engagement that was found to be positively correlated with positive outcome changes on the post-test. The researchers concluded that engagement was a significant implementation factor. The risk of bias assessment for this study did not raise any concerns.

Nazari et al., 2019. This was the third and final study rated as high priority in this area. It employed an RCT to test the effect of spaced practice on the mathematical performance of third and seventh graders (n = 213) in four German schools. Students were provided with a 90-minute introduction to a mathematical topic, derived from their curriculum. Thereafter, at class-level, students practiced in one of two conditions: in the massed condition, they worked on three practise sets in one day; in the spaced practice condition, students worked on one practice set per day for three consecutive days.

Key findings. Performance in two follow-up tests one and six weeks after the last practice set revealed a positive effect of spaced practice compared to massed practice in Grade 7. However, in Grade 3, a positive effect of spaced practice was supported by the data only in the test one week after the last practice set. The researchers concluded that spaced practice across several days improves mathematical performance of students in elementary and secondary school at least up to one week after the last practice set. The risk of bias assessment for this study raised some concerns about the randomisation process and potential deviations from the intended intervention.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section we focus on the study outcomes, summarised in the table below. Based on our review, summative grading of studies deemed as being both highly relevant to education practice and of high quality have been identified. These are marked with an asterisk in the table below.

Table B1.3: Spacing across days or lessons ('standard' spacing)—summary of evidence

Study	Focus	Population	Finding
		High Priority Studies	
*Feddern *et al.* (2018)	biology test scores	N = 829 Year 9 (13-14 years old Number of schools not reported. UK	***Neutral*** • 'Offline' spacing was not statistically significantly more effective than massed practice. However, the software mixed strategy condition (which included spacing alongside other strategies) produced significantly higher scores than both. Group means and statistical significance reported, but not SDs or effect sizes (see above for further details).
*Nazari et al. (2019)	maths	N = 141 3rd and 7th grade	***Positive*** • A) Both 1- and 6-week tests showed positive effect of spacing for Grade 7. With performance 6 weeks after the last practice set as dependent

		5 primary, 4 secondary schools, Germany	variable, the mean of the posterior distribution of distributed practice was 0.79 (95% CI = −0.35, 1.94). ***Neutral*** • B) For children in Grade 3, a significant positive effect of spacing found for 1-week test only. The posterior distribution of the effect of distributed practice on the performance 6 weeks after the last practice as compared with massed practice had a mean of 0.37 (95% CI = −0.50, 1.20).
*O'Hare et al. (2017)	science	N = 408 Year 9 and 10 students 12 secondary schools, England	***Positive*** • A) Combining 24-hour and 10-min spacing was most effective (particularly on long-answer questions) compared to BAU (*g* = 0.19, 95 % CI = 0.01, 0.36). • B) Engagement predicted more positive outcome change.
Larger Studies (pupil n > 500) (Medium Priority)			
Foot *et al* (2019)	critical thinking	N = 716 10-14 years 16 schools, 42 classrooms, US	***Positive (for fact learning)*** • Students in the spaced condition recalled facts than the massed condition. No difference between groups on critical thinking. Students in the spaced condition remembered more facts from the lessons (*d* = .21, 95 % CI = 0.07, 0.36) but showed no spacing advantage on the critical thinking measures where they had to explain their ratings in a paragraph.
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
Denton *et al*. (2011)	reading outcomes	N = 192 1st grade, at-risk readers 9 schools, US	***Neutral*** • Groups did not differ significantly on any of the 7 reading outcomes (*d* = 0.11, 0.18 ,0.11 ,0.10 ,0.02 ,0.13 ,0.21).
French et al. (1990)	volleyball skill	N = 139 1 high school, 4 classes, US	***Neutral*** • No difference on practice schedule between groups, for any volleyball skills.
Goossens *et al*. (2016)	vocabulary learning	N = 129 4th, 5th, 6th grade 2-3 classes per grade, 1 school Netherlands	***Neutral*** • A) No significant benefit of retrieval practice or spaced practice on either the cued-recall or multiple-choice tests. ***Negative*** • B) Some significant effects in unexpected direction: benefits of restudy in Grade 3, and short lag spacing in Grades 2 and 4.
Greving & Richter (2019)	biology text recall	N = 191 7th grade 3 schools, 8 classes, Germany	***Neutral*** • A) No overall effect of spacing on recall or comprehension. ***Negative*** • B) When using short interval between learning and test, massed group performed better on both measures.
Seabrook *et al*. (2005)	vocabulary learning and phonics	Three experiments N = 119, Years 1,3,6,9, 1 school N = 20, Year 2, 1 school N = 34, Year 1, 2 schools, UK	***Positive*** • A) More words were recalled when there was greater lag (i.e., more intervening items between to-be-learnt items) • B) Spaced presentations resulted in better performance than either massed or clustered presentations • C) Spaced teaching of phonics resulted in significantly better phonics improvement than clustered teaching
Svihla *et al*. (2018)	inquiry science learning	N = 139 9th-11th years of school 1 high school, 5 classes (2 chemistry, 3 earth science), US	***Neutral*** • No significant difference in either immediate or delayed test scores between the spaced and massed groups, η^2 = .01. In both conditions, students improved from post-test to delayed post-test, rather than forgetting information as is typical in studies of learning, η^2 = .33.
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Bloom *et al*. (1981)	second language vocabulary	N = 56 9th and 10th grade 1 school, 3 classes, US	***Positive*** • Performance was 35% better with spaced practice (M =15.04, SD = 3.78) than massed practice (M = 11.15, SD = 4.02) (*d* = 1.00, 95 % CI = 0.42,1.57).

Goossens *et al*. (2012)	vocabulary learning	N = 33 3rd grade 1 primary school, 2 classes, Netherlands	***Positive*** • A) After 1 week, words learnt through spaced learning (M = 55.96, SD = 26.24) were recalled more than words learnt via massed learning (M = 46.46, SD = 25.85) (*d* = .36) • B) After 5 weeks, this pattern remained but with smaller effect sizes: spaced learning (M = 49.49, SD = 27.13); massed learning: (M = 42.22, SD = 23.07) (*d* = .29)
Kupper-Tetzel *et al*. (2014)	EFL vocabulary	N = 65 6th grade 1 secondary school, 3 classes, Germany	***Positive*** • The optimal lag (i.e., spacing duration) depends on the retention interval: A) When vocabulary was tested after 7 days, the optimal lag was 1 day. B) When tested after 35 days, lags of both 1 and 10 days improved recall ($\eta^2 = 0.38$)
Namazian dost *et al*. (2019)	EFL vocabulary	N = 68 14-16 years 1 secondary school, 2 classes, Iran	***Positive*** • Students in spaced learning group showed greater improvements (M = 2.87, SD = 1.45), than the massed practice group (M = .15, SD = .36) (*d* = 2.57, 95 % CI = 1.93,3.21).
Nazari & Ebersback (2018)	maths	N = 44 10th and 11th grade 3 schools, 8 classes, Germany	***Negative*** • Small negative effect of distributed practice: lower test scores and higher drop-out compared with massed group. the students of the distributed practice condition were estimated to have a performance about 1 point (out of 15) lower than students of the massed practice condition (95% credible interval =−3.1 to 1.1).
Nazari & Ebersback (2019)	maths	N = 81 7th grade 4 schools Germany	***Positive*** • A) 2-week post-test: small effect of practice condition (*d* = .12) • B) 6-week post-test: positive effect of spaced practice (*d* = .33, 95 % CI = 0.77, 0.11) (Given the theory for spacing, we have taken this as the main effect). • C) Exploratory analyses: students in the medium performance range benefitted the most
Peterson-Brown *et al*. (2019)	Maths vocabulary	N = 62 3rd and 4th grade 4 elementary schools, US Quasi-expt (assignment at individual level)	***Positive*** • A) Maths vocabulary scores better in fixed and expanded conditions (M = 5.56, SD = 1.86) compared with massed condition (M = 4.36, SD = 1.94) (*d* = 0.64, 95 % CI = 0.10, 1.17) • B) No differences in scores between fixed interval and expanded interval spaced practice
Sobel *et al*. (2011)	vocabulary learning	N = 39 5th grade 1 middle school, 2 classes, US	***Positive*** • Students recalled more words correctly via spaced learning (M = 20.8, SE = 4.3) than massed learning (M = 7.5, EM = 2.0) (*d* = 0.48).

* High priority study, identified for in-depth analysis; ^ = study included for more than one strategy.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in Table B1.4 below.

Table B1.4: Spacing across days or lessons ('standard' spacing)—quality of evidence assessment (based on the GRADE approach)

Strategy	'Standard' spaces—across lessons or days
Number of studies	There are 18 studies in this area of which three were rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	Seventeen studies are randomised experiments; there was one quasi-experiment (Peterson-Brown et al., 2019).

Risk of bias	Our risk of bias assessments on the high-quality papers did not identify concerns with two of the studies but some concerns with randomisation and fidelity to treatment with one. We judge, therefore, there to be at least two strong studies in this area.
Inconsistency	**Result consistency**. Results in this area suggested a range of results from small negative to moderate positive effect sizes. Most results were around zero to an effect size of 0.3.
Indirectness	**Practice heterogeneity**. Studies in this analysis area were judged to be testing a highly similar (homogenous) practice of spacing across lessons—spanning several days and in some cases several weeks. **Population, measure, and outcome heterogeneity**. There were appreciable differences in the learning outcomes and the practices and learning approaches. The high priority studies were focused on science and maths. Studies of reading and vocabulary-learning in the medium-sized study group provided more mixed results. There is therefore a concern about subject applicability beyond maths and science, with only Foot et al. (2019) providing a trustyworthy positive result out of these areas (in critical thinking). **Outcome measures**. Many of the interventions relied on researcher-developed and focus learning content-specific vocabulary tests. Tests of learning outcomes are—for all but a very small number of studies—limited to simple recall items on researcher-designed and content-aligned tests. **Design and delivery**. Overall, ecological validity in the section was moderate, with studies focusing on testing spaced practice principles in real classrooms, but often not delivered by regular classroom teachers or as part of the regular curriculum.
Imprecision	**Group sizes**. Sample sizes (at pupil level) varied: four studies had a pupil n of 50 or less, nine had 51 to 150, and five had 151 or more. Many studies randomised at a class or school level. Where studies report effect sizes, these tended to be in the vicinity of *d/g* = 0.1–0.3. We judge O'Hare et al. (2017) to be the highest quality and most precise study in this group. This study reported an effect, g, of 0.19 (95 % CI: 0.01, 0.36).
Publication bias	A high proportion of positive results for smaller studies relative to medium and large studies suggests publication bias is present for smaller studies and estimates will be more trustworthy when based on medium, large, and high priority studies.
Other considerations	The larger, medium-priority study, Foot et al. (2019), while it was focused on critical thinking, found an effect on fact learning, but not critical thinking.
Overall confidence	**Low (++)** Our confidence in the effect estimate is limited: the true effect may be substantially different from our estimate.
Confidence reasons	This low confidence is based on the following issues: - the tendency towards publication bias for smaller studies; - a limited range of subjects (maths, science, or critical thinking or fact learning) and a small number of high priority and larger studies; - few trustworthy studies reporting effect sizes on which to base effect estimate; and - test outcomes limited to mostly research simple recall items on researcher-designed and content-aligned tests.

Summary of findings for this strategy

Main finding. Overall, evidence suggests that spaced practice has a small positive effect on learning compared with massed practice in maths and science. Results in other areas, such as reading and vocabulary, suggest little to no effect, with some negative results.

Estimated impact. We judge O'Hare et al. (2017) to be the highest quality and most precise study in this group. This study reported an effect, g, of 0.19 (95 % CI: 0.01, 0.36) for science outcomes. Similar estimates are obtained by Foot et al. (2019)—d = 0.21 (95 % CI: 0.07, 0.36) for critical thinking—and, although none were statistically significant, Denton et al. (2011): seven effect sizes, d = 0.11, 0.18, 0.11, 0.10, 0.02, 0.13, and 0.21.

Confidence in impact estimate. With mixed evidence in this area, and considerable areas of inconsistency and study limitations, we have rated this finding as having low security.

Heterogeneity. Given the limited evidence, we did not conduct further heterogeneity analysis.

Strategy 2: 'Short' spacing

Concise definition

'Short' spacing involves spacing learning within a single lesson period, usually separated by short intervals where students complete an unrelated task.

Full definition and description

'Short' spacing involves spacing learning within a single lesson period, usually separated by short intervals where students complete an unrelated task. This interval, often referred to as the inter-study interval (ISI) or 'distraction' task, is typically around 10 to 15 minutes in length. In instances where the ISI is a related topic or task, this may be an example of interleaving (Strategy 3).

Selected examples

Examples of this strategy from our database include:

- O'Hare et al. (2017) report their trial of the SMART Spaces programme. This compared three treatment conditions (as summarised in the table below) and one control condition (receiving slides but no spacing protocol).

	Version 1 10-minute spacing	Version 2 24-hour spacing	Version 3 Mixed
Day 1	▪ 12 minutes of chemistry ▪ 10-minute 'space' ▪ 12 minutes of chemistry repeated ▪ 10-minute 'space' ▪ 12 minutes of chemistry repeated	▪ 12 minutes of chemistry ▪ 12 minutes of physics ▪ 12 minutes of biology ▪ 20 minutes of 'space' at end	▪ 12 minutes of chemistry ▪ 10 minutes of 'space' ▪ 12 minutes of physics ▪ 10 minutes of 'space' ▪ 12 minutes of biology
Day 2	As day 1 but for physics	As day 1	As day 1
Day 3	As day 1 but for biology	As day 1	As day 1

- Spaced learning in Kelley and Whatson (2013) consisted of 'three intensive instruction elements of the same content with minor variations each lasting 20 min or less (stimuli), spaced by two distractor activities of ten minutes (spaces without the stimuli)', which they compared to (remarkably) four months of regular teaching in biology.

Evidence for this approach

There were two studies in this area, both graded as high. The results for the risk of bias analysis for these are summarised in Table B1.5, below.

One of these, Churches et al. (2020), was a meta-analysis of 34 co-ordinated, small, teacher-led RCTs, of which six were focused on spaced learning. For studies across all cognitive science areas, teachers were provided with an RCT design day and pre-reading material about RCT design and cognitive science concepts. Teachers then designed and led their own RCTs. Curriculum subjects spanned mathematics (times tables, problem solving), English (vocabulary, spelling), science, history, and geography. Trial length varied from a single lesson to 42 days. Of the 34 RCTs, six focused on spaced learning. Those six studies were not reported in detail individually, but details of their topic area, n, effect size, and an analysis of their robustness are provided. We reproduce an overview of the

individual studies below. All studies are reported as using ten-minute intervals for spacing. They were all from the same author, reported in two publications.

Table B1.5: Summary of Churches et al. (2020) meta-analysis—spaced learning trials

Author	Year group	Subject	n	Effect size (d)	Jadad score for robustness (0–5)[5]
Bryant-Khachy (2018b)	Y5	History	54	0.85	3
Bryant-Khachy (2018b)	Y4	History	56	0.61	3
Bryant-Khachy (2018a)	Y2	Geography	60	0.43	3
Bryant-Khachy (2018b)	Y6	History	57	0.28	3
Bryant-Khachy (2018b)	Y3	History	223	0.12	3
Bryant-Khachy (2018a)	Y1	Geography	50	0.04	3

Key findings. Overall, these studies suggest a positive impact of within-lesson spacing. All studies are in either geography or history, and from the same author. All are for primary school age children. Most have a sample size of 50–60 apart from one that has 223. There are no details about why the effect sizes might vary, given the ostensibly highly similar conditions. This study's risk of bias assessment raised some concerns with the randomisation process and about deviations from the intended intervention. Note that the underlying studies were not accessible and full details were not provided. Our assessment of risk of bias, raising some concerns, reflects some gaps in the information provided as well as concerns based on the details provided.

The second study testing short, within-lesson spaces was Kelley et al. (2013). This study investigated the impact of spaced learning on biology test scores. This was a large RCT with group allocations randomised at the individual level. Students were 13 to 15 years old, from one school in England (n = 440). The publication presents results from three experiments (referred to as 'conditions'), as follows (p.5):

> *In Condition 1 students aged 13–15 were randomly assigned to experimental (n = 46) or control groups (n = 127). Condition 1 was constructed in part to restrict any learning other than through Spaced Learning, and ensuring [short term memory, STM] was minimized or eliminated by having five days between the Spaced Learning session and the test.*
>
> *In Condition 2 students aged 14–15 were in ability-matched groups from the beginning of the academic year, and one was randomly assigned to the experimental condition (n = 21) and controls were in similar-sized groups (n = 131) […] In Condition 2, the normal educational context was preserved as far as reasonably possible with all students with their own group and teacher, and having completed the first Biology course before taking the second Biology course. Then, for the second Biology course, the experimental groups experienced the Spaced Learning session and were tested. Controls were taught over four months and were tested.*
>
> *In Condition 3, experimental subjects aged 14–15 were taught the first Biology and Physics courses in the same teaching groups and then were tested (n = 115). Condition 3 was designed to test any impact of Spaced Learning after normal teaching of a course,*

[5] The Jadad scale, sometimes known as Jadad scoring or the Oxford quality scoring system, is a procedure to independently assess the methodological quality of a clinical trial with score of between zero (very poor) and five (rigorous).

remove the novelty of Spaced Learning without teaching, and enable more direct comparisons with another subject (Physics) and students in other schools […] At the end of the Physics course all students had an intensive one-hour review of all course content days before the examination, as is common practice in English schools. In contrast, at the end of the Biology course, this intensive review was replaced by a single, spaced Learning session of the same duration. In studies 1 and 2, biology curriculum content was delivered in a single, compressed spaced learning session of 60-mins. The same content was repeated 3 times with minor variations, with 10-min spaces in between. This was compared with a control condition which had 4 months (23 hours) of traditional biology teaching.

(Kelley et al.*, 2013, p.5)*

The general procedure was that after training on spacing and memory, teachers taught using their own materials. Teachers chose their own (physical) distractor activities (such as basketball practice, juggling, and clay modelling) for the within-lesson spaces. A multiple-choice biology test was used as the outcome measure, based on GCSE biology questions.

Key findings. In summary, the findings of Kelley et al. (2013) were that scores did not differ significantly when content was delivered in a spaced 60-minute session (treatment) versus the four months (control). However, it is necessary despite this null result to note the rate of learning: given the amount of time spent on learning in each group, spaced learning groups learned considerably faster per hour of instruction. Moreover, as tested in condition three, when a traditional biology end-of-course review was replaced by a spaced equivalent, test scores were significantly higher, an increase of 7.6%. The risk of bias assessment for this study raised some concerns with the randomisation process, deviations from the intended intervention, and selection of reported results.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section we focus on the study outcomes, summarised in the Table B1.6. Studies identified as high relevance and quality have been marked with an asterisk.

Table B1.6: Spacing within lessons—summary of evidence

Study	Focus	Population	Finding
High Priority Studies			
*Churches *et al.* (2020)	History and Geography	6 studies n = 54, 56, 60, 57, 223, 50. Year 1 – 6 students, England	***Positive*** • Effect sizes were positive for all 6 spaced practice trials but only one was statistically significant. Year 5 History (*d* = .85, p<.01), Y4 History (*d* = 0.61), Y2 Geography (*d* = 0.43), Y6 History (*d* = 0.28),Y3 History (*d* = 0.12), Y1 Geography (*d* = 0.04)
*Kelley *et al.* (2013)	Science Biology	n = 440 13-15 years old 1 school, England	***Positive*** • Scores did not differ significantly when content was delivered in a spaced 60-min session (treatment) or 4 months (control) • However, spaced learning groups learned significantly faster, demonstrated by % performance gain per hour of instruction • When a traditional Biology end-of-course review was replaced by a spaced equivalent, test scores were significantly higher (+7.6%, *d* =0.53, 95 % CI = 0.33,0.72)

* High priority study identified for in-depth analysis.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in the table below.

Table B1.7: Spacing within lessons—quality of evidence assessment (based on the GRADE approach

Strategy	'Short' spaced practice (within lessons)
Number of studies	There were two studies in this area; both were rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment. One of these, Churches et al. (2020) was a meta-analysis of 34 small, teacher-led RCTs of which six were focused on spaced learning.
Design	One study is an RCT, the other a meta-analysis of six RCTs.
Risk of bias	Our risk of bias assessments identified concerns with both studies in this area. A key issue was that insufficient details were published to confidently assess risk. Churches et al. (2020) conducted a Jadad robustness analysis, grading all six studies as '3' on a 0–5-point scale, but we were not able to verify the details of this or conduct our own analysis of the underlying studies. Risk of bias analysis of Kelley et al. (2013) raised some concerns with the randomisation process, deviations from the intended intervention, and selection of reported results.
Inconsistency	**Result consistency**. There is large variation in the results, even with the single cluster of studies in Churches et al. (2020). Estimates for individual RCTs ranged from around zero to $d = 0.85$.
Indirectness	**Practice heterogeneity**. There was not enough detail reported in Churches et al. (2020) to assess this point. The single authorship and subjects might suggest that the Churches' meta-analysis studies were relatively homogenous; few details were provided. However, the procedure is likely to differ markedly from that in Kelley et al. (2013). **Population, measure, and outcome heterogeneity**. Both studies were conducted in a single school in England. There is insufficient variation in the population to allow a definitive test. Outcomes were science, geography, and history, with a good range of ages.
Imprecision	**Group sizes**. Kelley et al. (2013) and one study within Churches et al. (2020) provided a good sample size for providing a potentially precise estimate. The other five studies reported were small in scale. Kelley et al. (2013) estimated a larger $d = 0.53$ (95 % CI: 0.33, 0.72). Effect sizes in Churches for six spaced practice trials were positive but only one was statistically significant. Year 5 history: $d = 0.85$, $p < 0.01$; Y4 history: $d = 0.61$; Y2 geography: $d = 0.43$; Y6 history: $d = 0.28$; Y3 history: $d = 0.12$; and Y1 geography: $d = 0.04$.
Publication bias	There are no suggestions of publication bias. Given the co-ordinated nature of the trials in Churches et al. (2020) we judge the probability of publication bias to be low.
Other considerations (including upgrading)	With one plus six studies in total in the area (one RCT and one meta-analysis reporting six studies), we judged this area as being suitable for an evidence assessment. On further analysis, however, we judge there to be three distinct studies reported: - Bryant-Khachy (2018a); - Bryant-Khachy (2018b); and - Kelley et al. (2013). While this initial evidence is promising, it is too early to reach an evidence-based judgement on the effectiveness of this strategy without replication of these studies.
Overall confidence	**Very low (+)** We have very little confidence in the effect estimate. The true effect is likely to be substantially different from the estimate of the effect.
Confidence reasons	This very low confidence is based on the following issues: - the low number of studies: at most there were seven RCTs reported; six, however, related to trials by a single author that amounted to two distinct studies; - as Churches et al. (2020) was a meta-analysis and the underlying studies were not publicly available, we were not able to conduct full analysis of the underlying studies; and - there was considerable unexplained variation in effects in Churches et al. (2020).

Summary of findings for this strategy

Main finding. There is insufficient evidence to assess the effectiveness of this strategy. Indicative evidence suggests that 'short' or within-lesson spaced practice has a positive effect on learning compared with massed practice and facilitates the time-efficient learning of content.

Estimated impact. Effect size estimates were *d* = 0.53 (95 % CI: 0.33, 0.72) for Kelley et al. (2013) and between *d* = 0.04 and 0.85 in Churches et al. (2020), six results, with only the largest statistically significant. This suggests moderate potential effects. This is indicative only, however, and is based on too few studies (in scale and authorship) to form a confident judgement.

Confidence in impact estimate. This finding receives a very low confidence rating. The primary weaknesses of the studies in this area are the limited number of tests and, for individual RCTs, the large range of effect size estimates with no evidence that explains the large variation.

Heterogeneity. The evidence was insufficient to assess heterogeneity in this area.

Spaced practice—overall evidence summary and conclusions

Summary of results

In this section, we have reviewed 27 studies focused on spaced practice. We identified two strategies for which we potentially had sufficient evidence to assess effectiveness. Our results for these are summarised in Table B1.8.

Table B1.8: Spaced learning—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
'Standard' spaces, across lessons or days	Eighteen, of which three were graded as high priority.[1]	The overall evidence suggests that spaced practice has a small but positive effect on learning compared with massed practice.	There was a good age range (6 to 17) represented. There were a range of subjects, including literacy, maths, science and PE—although this was limited for larger and high priority studies (to maths, science, and critical thinking).	Low (++)
'Short' spaces, within lessons	Two, of which both were graded as high priority;[1] one a meta-analysis of six small-scale trials.	The evidence suggests a positive effect on learning compared with massed practice and that it might be a way of learning content in a highly time-efficient manner.	Outcomes were science, geography, and history, with a good range of ages. Although, there are too few studies here to reach a judgement about applicability.	Very Low (+)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Conclusions about strategies in this area

Spaced practice

Our headline conclusions in this area are:

- Spaced practice is *potentially* highly relevant across the U.K. education system, for all learners and subjects. The spacing of learning is a fundamental aspect of curriculum and lesson design. Longer spaces affect curriculum design, within and across school years; shorter spacing is highly relevant to lesson planning and pedagogy; standard spacing is potentially relevant to both.

- The results suggest a small but positive effect for spaced practice (*d* = 0.2 based on the highest precision study and other results in the 0.1–0.2 range).
- The high priority and largest studies represent a more limited range of studies (maths, science, and critical thinking), with a suggestion that effects in other areas are lower or less consistent.
- There was more evidence for 'standard' spacing (across lessons and days) than within-lesson spacing. There was indicative evidence for the latter, and it follows from the theory, but there is too little evidence to reach a firm judgement.

The implications of the evidence presented above is that spacing is a plausible strategy for promoting additional learning. Still, there is large variation in the practice, and the evidence-base is currently relatively limited, even for assessing whether spacing is, in general, effective. More robust research on the overall effect of spacing and moderating factors is needed before firmer conclusions and more confident recommendations can be made about whether to space learning and how to do it effectively.

While these results do not suggest a large impact of spacing, it was one strategy area that may be possible to implement at scale through building spacing into units or schemes of work at the planning stage. However, we also note that spacing may create further demands on an already crowded curriculum. This issue is one we explore further in our discussions and questions section for spaced practice.

Evidence-informed discussion and questions

Principles and moderating factors

Spacing as a 'desirable difficulty'

How does the spacing effect work?

Educators, policymakers, and researchers may well want to know why spacing out practice or learning might be more effective or efficient for consolidating learning than massed practice (Smolen et al., 2016). The relationship between brain activity and memory seems key here. Dehaene (2020) explains that brain imaging studies demonstrate that massed practice generally correlates with decreased brain activity, compared to spaced learning. He suggests this is because spacing 'seems to create an effect of "desirable difficulty" by prohibiting simple storage in working memory rather than long-term memory, and thus forcing the relevant circuits to work more' (p.218). Similarly, Callan et al. (2010) link spacing with *greater* levels of brain activity linked to maintenance rehearsal, compared to massed practice in the context of vocabulary learning. This is a point echoed elsewhere, with Brown, Roediger and McDaniel (2014) explaining that massed practice can be deceptive: it feels easier and feels like one knows more. Spaced practice feels harder, and less certain, but this is a desirable difficulty as it forces us to make greater use of our long-term memory. One of the teachers we interviewed felt pupils, particularly low attaining pupils, 'disliked' spaced practice, favouring massed practice 'because completing loads of work probably gives them a feeling of success that slower, spaced, and retrieval tasks don't'. If this observation is correct, it could have important implications for attempts to design learning that both maximises learning through spaced practice and ensures attainment gaps are not exacerbated.

Decreased brain activity, however, does not necessarily mean less processing; it may just reflect less *efficient* processing. The role of attention might be a factor when comparing spaced and massed

practice: massed practice is likely to place greater demands on attention resulting in less efficient learning. Within our broader evidence, there was one study, Chen et al. (2018), that concluded that the spacing effect might be 'directly attributed to working memory resource depletion'. However, they acknowledge that their results do not eliminate the possibility of other causes under different conditions. Implications of this study would suggest higher cognitive effort is required for *massed* practice.

In the literature we reviewed, therefore, there are two explanations for a spaced practice effect. The first, emphasising the beneficial additional demand on long-term memory for spacing, a 'desirable difficulty', and the second emphasising the negative impact of additional demand on working memory, and its depletion. This tension within the evidence may have important implication for designing interventions in real classrooms. The type of practice implemented in the classroom in our wider evidence is typically a form of retrieval (from long-term memory), suggesting that the beneficial additional demand is the primary factor at play in most instances of classroom spaced practice.

Learner characteristics and practice accuracy

Does learners' prior knowledge or practice accuracy affect the spacing effect? How?

Educators, policymakers, and researchers may also want to know how prior learning inter-relates with spacing. Another study within our wider evidence was Kasprowicz et al. (2019), who investigated the impact of practice distribution in combination with language analytic capacities in primary classrooms; 113 beginner-level English learners of French (aged 8 to 11) were included in a four-week pre-experimental phase followed by four lessons of introduction to core vocabulary. After this, classes were assigned to a control group that received three sessions of 60 mins with seven days in between and a group that received six sessions of 30 minutes with 3.5 days in between. Their results showed that there were minimal differences between the treatment and control groups, but that the shorter spacing may have been slightly better for the target group. They also found that learning was moderated by the practice accuracy of both groups and the language analytical ability for the 3.5-day condition. Although we do not have further information in this area, learner prior knowledge and practice accuracy are plausible moderating factors for the spacing effect.

Variation in the practice or teaching and learning context.

This section explores some of the variation in either the practice of spaced learning or the teaching and learning contexts in which it has been applied.

Spacing Intervals

What is the optimal spacing interval and what does it depend on?

One key variable for spaced practice is the length of the interval between learning or practice sessions (the 'inter-study interval'). In the evidence we report above, we distinguish between standard and short spaces, though the latter did not have sufficient evidence for an assessment of effectiveness. For standard spacing, we recognise that there is considerable variation in the length and type of teaching and learning activities in a session, the interval between sessions, the number of sessions, and the interval to a final test. These are likely to influence the extent of learning and its measurement.

Even in our assessment of standard spacing, where there were 18 studies, there were not enough comparable studies for more granular analysis of interval size as a moderating factor.

Consulting the wider evidence and literature, the prevailing view on an optimal interval is that it depends on the desired duration for memory retention. Dehaene (2020, p.218), referring to earlier work (Kang et al., 2014), discusses this principle. He suggests 24 hours may have benefits as this allows the student to sleep between sessions; sleep is known to be important for memory consolidation (Buzsák, 1998). Dehaene also suggests a rule that intervals should be 20% of desired memory duration. So, for example, one should rehearse after two months to retain for ten months. He also mentions the strategy of starting with multiple close spaces and then gradually increasing intervals. What is interesting about this way of looking at memory is how the decision is framed in terms of future retrieval and memory strength rather than current or past strength.

> *'Indeed, we may have been wrong about memory: it is not a system which is orientated towards the past, but one whose role is to send data to the future, so that we may later access it. By repeating the same information several times, at long intervals, we help our brain convince itself that this information is valuable enough to be delivered to our future self.'*
>
> *(Dehaene, 2020, p.219)*

This account chimes with the wider account evidence in the practice and research literature (for example, Cepeda et al., 2008). This suggests that memory gradually declines over time, and that spacing intervals can be designed with reference to Ebbinghaus' famous 'forgetting curve' in mind (also see Küpper-Tetzel et al., 2014).

Without sufficient applied evidence in this area, we are not in a position to endorse any decision rule on the optimal spacing interval. Further applied evidence in this area would be of great value for teachers implementing spaced learning in their classrooms.

Looking at our own wider evidence, there were some examples of studies looking at spacing over a period of months rather than hours, days, or weeks. Collins and White (2011) and Collins et al. (1999) compared spaced practice over five to ten months. In these studies, all groups made substantial progress. In the former, there was no clear difference between either group; for the latter, it was the massed condition group that made the most progress. Collins et al. (1999) suggest that learners may reach a plateau whereby more hours may not bring substantial additional learning. In both studies, they also found that both groups did make considerable progress, and that there were varying moderating factors such as different use of homework and the distributed condition group having to study for their end-of-year exams at the same time as the interventions. To a large extent, the lack of—and mixed—evidence for longer spacing intervals is likely to reflect the difficulties of assessing this within controlled conditions.

Feedback, personalisation, and assessment for learning

There were two studies in our wider evidence examining interaction and personalisation around spaced learning or spaced practice sessions. The effectiveness of spacing is likely to be influenced by practice schedules being tailored to prior learning or responding to practice accuracy and feedback on, or additional teaching, of weaker areas.

To what extent should feedback be built into spaced practice sessions? Does this depend on the practice accuracy or learning content?

Codding et al. (2019) looked at 'opportunity to respond' as another possible moderating factor for spaced practice. They compared fact fluency in maths in the U.S. context through a pilot study using four treatment conditions combining massed and distributed practice with high and low opportunity to respond in a class-wide application of 'cover-copy-compare'. Their findings suggest that the opportunity to respond has a more significant impact on learning than the spacing effect, and that insufficient opportunity to respond has significant negative implications. We note that adding in feedback potentially introduces informational and motivational influences that could be seen as an additional learning strategy in its own right, rather than as a moderator of the effect of spacing *per se*. From the perspective of the science, there is value in isolating spacing-specific moderating and mediating factors. From the perspective of practitioners, identifying feasible ways to combine strategies may make this coupling of feedback and spacing a benefit rather than a problem.

Lindsey et al. (2014) investigated the potential of using a computer programme to schedule and personalise spaced practice. They compared three conditions in a study of eighth-grade Spanish foreign-language instruction where students studied ten chapters of material with a week in between each. After each session, the students reviewed the material for 20 to 30 minutes with a Colorado Optimized Language Tutor. The material they reviewed was selected by a scheduler following three different formats: the 'massed scheduler' selected material from a chapter that students had least recently studied; the 'generic spaced scheduler' selected a chapter previously studied for review based on what was considered optimal for a range of students and materials; and the 'personalized spaced scheduler' used a latent-state Bayesian model to predict what specific material a particular student would most benefit from reviewing. Findings from the study suggest that a one-size-fits-all variety of review or retrieval is significantly less effective than a personalized one.

Can—and should—spaced learning be differentiated or personalised? What is the role of assessment for learning and ongoing assessment in the successful implementation of spaced practice?

Variation and transfer

Should learning content and approach be varied when spacing? What is the effect of this on outcomes? Does variation promote transfer?

One final point raised in our analysis was the potential of spacing practice to improve learning transfer. It is thought that spacing might enhance the transfer of learning across problems and contexts (Carpenter and Agarwal, 2020; Gluckman et al., 2014) . Related, Tibke (2019, p.78) suggests that 'retrieval might be improved by replicating the conditions surrounding the initial coding' (also see Medina, 2008). In our wider evidence (for example, Sobel et al., 2011, p.765; Peterson-Brown et al., 2019), there is a discussion of contextual variability and whether this supports retrieval and is beneficial for deeper learning and transfer. Overall, the evidence is not clear on when, how, and if variation in learning during spaced learning is advantageous. Our intuition suggests that when learning is new and memory strength is weak, minimising variation will promote recall, but that variation will become increasingly valuable as pupil knowledge and familiarity increases, and that variation might be used to promote transfer. This is a question we return to in the next section when examining the evidence for interleaved practice.

Gradually and deliberately changing the learning task being spaced was something discussed by several teachers in our practice review data. One teacher explained that when spacing learning they 'gradually work up to things which are more challenging, gradually deeper knowledge, so we start with

the basic knowledge and then we gradually move up to the evaluation application skills with them in terms of deeper processing' (Interviewee 2). Another teacher discussed their daily morning routine of spaced retrieval practice with approximately 30 maths questions from different maths topics. Subtle but deliberate changes were made in the questions each day to extend as well as consolidate learning.

Implementation

What kind of learning should be spaced? What approaches are there?

Thus far, we have treated spaced learning as a singular strategy and largely used the term synonymously with spaced practice. The evidence did not enable deeper exploration of the questions around the level of familiarity with material (for example, spacing new learning versus spaced revision of previous learning) or the specific approaches teachers use to space. The practice review data, however, did provide examples of teacher practice that we briefly summarise below.

Overall, teachers mostly described spaced practice as a form of retrieval practice with low-stakes, short-answer quizzes or tests repeated over time with deliberate spacing out of a topic. Discussion of within-lesson spacing was rare, although there were examples, including one teacher (Interviewee 3) who described a lesson in which students had short inputs and practice interspersed with 'unconnected activity so that information doesn't stay in working memory'. In terms of specific classroom activities, examples included (in respondents' own words):

- 'a five minutes exercise at the beginning of every lesson, or before the break period';
- 'three multiple-choice questions at the start of each lesson (use as a starter activity)';
- 'creating recall grids that cover clusters of topics, with extension questions that link some of the topics together';
- 'synoptic exam papers and questions, starter recall quizzes or activities';
- 'a "spiral curriculum" model where we return to previous knowledge throughout';
- 'gap tasks in order to scaffold learning and challenge all learning groups';
- 'for revision, one lesson in a fortnight on a small aspect of a topic for revision';
- 'regular starter quizzes with KS4 and five physics students to improve their retention of key facts/equations';
- 'the start of every lesson … five questions every lesson: one from any time, two from last year, three from last topic, four from last week, five from last lesson';
- 'students have a pack of starters that we work our way through based on work from previous topics; these could be diagrams to label, questions to answer, gap fills to complete that are completed and self-marked by the students to encourage hard thinking and resilience—the expectation is that students will at least have a go';
- '[spaced] structured homework with KS3; I did this as they needed it structured for them, whereas older pupils wanted more freedom'; and
- 'using Quizlet or Quizizz'.

Teachers tended to discuss spaced practice as something done 'little and often' or in 'short intense' practice sessions before moving on. Spaced learning often involves a previous topic being introduced with a current one. Spaced practice was often described as a form of revision, described using terms such as 'throwback' or 'interrupt' lessons. Many classes had become used to having revision lessons and activities and the practice of doing 'revision lessons as we go'.

There was some discussion about which subject areas are suitable for spaced learning. Maths was mentioned on several occasions as a subject that lends itself well to spaced learning due to its 'spinal curriculum' and the tendency to 'reuse ideas and build on them towards mastery'. A range of other

curriculum subjects were also discussed but the common element was often about emphasising core concepts or ideas that were important to the subject and were shared across topic areas. For example, one teacher described practice in English in their schools as follows:

> *I think spaced practice within English is standard—as is retrieval and interleaving. As a subject, you need to make connections between texts, recognise tropes, genre etc. and compare and contrast. This is pretty much the essence of the subject. Across a three-year KS3 curriculum, pupils will encounter the same concepts again and again in different texts and have to do the same sorts of things again and again.*
>
> *Questionnaire respondent*

Spaced practice seems particularly relevant to subjects where students are required to draw on their knowledge of mixed topics, apply mathematical strategies, talk at length across topics in modern foreign languages (MFL), or complete synoptic units, for example. There were teachers with reservations about the application of spaced practice across *all* curriculum areas, however. One interview participant linked ideas about knowledge and skills-based curricular and particular subjects as follows:

> *We definitely follow what you call a more knowledge-based curriculum. Spaced learning is perhaps harder in more skills-based subjects, with art with music. With DT we deliver a lot of content around knowledge, for example, going into the different stages of designing products, so the children are very familiar with that process but, for example, in art, at the end of the day, it is still about whether they are able to do it, have that artistic skill. And that's harder, every week they have art and there is spacing from the previous week and building on it. But I think it is harder.*
>
> *Interviewee 11*

What demands on curriculum planning and timetabling are made by spacing? Should spacing be prioritised for core curriculum content?

Spacing practice has an organisational component for teachers to manage—how to organise lesson time and sequence a curriculum that allows for the desired spacing in learning and practice intervals. Strategies such as setting homework and using low-stakes quizzes on earlier content have proved popular to make spacing work in practice. Several questionnaire respondents discussed how they planned spacing into their schemes of working and that this 'only takes a small amount of planning, but the results are great'. However, many teachers spoke to us about some of the difficulties of applying spaced practice. The challenges roughly grouped into the following, using direct quotes from our survey or interview data:

Crowded curriculum and time

- The curriculum is crowded, with too much to be 'covered within a specific time frame', so taking lesson time away to revisit previous content 'is a luxury that is rare'.
- 'There is no time to recap or interweave previously taught material.'
- 'Exam syllabus restrictions can limit opportunities to develop spaced practice and interleaving.'
- '[Spaced practice is] easy over a few years but in a GCSE or a level course it's tricky to have sufficient teaching time.'

- ‘Spaced practical, and to a degree retrieval practice, both need time, the first for planning and the second in terms of lesson time.’

School timetables and curriculum planning

- ‘Talking to other teachers, one concern is common: that the schemes people are using doesn’t allow for [spaced learning]. The resources are very well prepared, question and answer sheets. Teachers tend to pick these up and just deliver’ (Interviewee 6).
- ‘School timetabling tends to block subjects, making spacing more difficult.’
- ‘[It’s] more an organisational challenge than a reflection of implicit difficulty in implementation.’
- ‘Spaced practice [is] more difficult to implement, as not a whole-school policy [so] not well-recognised by leadership [and] seen as an unknown.’
- ‘Spaced practice has been tricky to plan whole school through a two-year rolling programme.’
- ‘Spacing the scheme of work makes it difficult—[I] need [the] whole department to adjust.’
- ‘I think the biggest difference that I notice in the way that you deliver primary versus secondary is I could be constantly spacing my practice across the whole curriculum in primary—and so I didn't need to structure the space practice and retrieval as rigidly’ (Interviewee 10).

Another practical consideration touched on in the wider literature is that studies tend to suggest that benefits of spaced practice are likely to arise after a delayed, rather than immediate, test (Weinstein, Madan and Sumeracki, 2018). This may have implications for classroom assessment when assessing learning over time.

Final thoughts on this strategy area

In our systematic review of classroom trials, we concluded that spacing is a plausible strategy for promoting additional learning. However, there is large variation in the practice and the evidence-base is currently relatively limited, even for assessing whether spacing is, in general, effective. The results suggested a small but positive effect for spaced practice (d = 0.1–0.2 based on the highest-precision studies) and our overall confidence in this effect was rated as low.

In our discussion, we have examined the wider evidence and literature and posed numerous questions about spacing. For example, we have examined how spacing works, possible moderating factors such as pupil prior knowledge, and possible connections with feedback. In general, the literature identified plausible scientific and professional interpretations, but little that was grounded in evidence. The question of spacing interval is particularly fundamental to the strategy and would be a fruitful area for further research to examine in applied settings. As we noted earlier, our main results do not suggest that spacing has a large impact but rather that spacing is one strategy area that may be possible to implement at scale through building spacing into units of work or schemes of work at the planning stage. This, and the positive but tentative results, suggest promise. However, as per the discussion above, spacing may create further demands on an already crowded curriculum, raising practical as well as pedagogical questions for successful implementation.

B2. Interleaving

Overview of area

Definitions

This section focuses on studies of interleaved learning. When learning tasks are interleaved, they are, inevitably, also spaced. This can make the two hard to distinguish from each other (Agarwal and Bain, 2019), both practically and conceptually. The previous section focused on the spacing of learning or practice. Interleaving is similar but distinct from spacing. Interleaving consists of sequencing learning tasks so that *similar* items are interspersed with slightly (but not completely) different types of items rather than being presented consecutively (Rohrer et al., 2019). In spaced practice, on the other hand, spaces are usually filled with *unrelated* activities or the learning of unrelated topics.

Interleaving also forms an important link to another area of this review: Working with Schemas. In that section, one of the cognitive science strategies is the use of variation and comparison to develop thinking and knowledge (schemas). As we discuss in detail in the dedicated section on this, comparisons, analogy, and presenting cognitively conflicting information is thought to benefit learning. Interleaving apparently forms the intersection between these two ideas: spaced learning combined with a form of comparison. This is one plausible explanation for why interleaving offers 'added value' over and above the potential benefits of spaced learning. We consider this at greater length in the Discussion and Questions section after reviewing the evidence for interleaving.

Overview of the evidence-base

Table B2.1: Interleaving studies—overview of priority ratings

Priority Level	Overall rating	Ecological validity	Relevance and definition for focus CS practices	Added value to evidence-base
High	6	3	9	5
Medium	6	9	3	11
Low	4	4	4	0

The review database contained 16 studies in the interleaving category. Of these, 12 were graded as being of sufficient ecological validity, relevance, and value for inclusion within this analysis of the evidence (high and medium). Six studies scored highly across these criteria and were identified as *potentially* providing strong evidence in this area (high).

When studies reported a specific focus on interleaving, they could confidently be categorised in this section. Of the 12 papers in this analysis, nine were rated as 'high' in terms of relevance and adherence to the definition. The limitation of this was that outside of studies explicitly reporting a focus on interleaving, we had many studies that drew on the same general principles. As we discuss above, there are strong connections to spaced practice and ideas around the comparison. Arguably, there is a larger evidence-base testing the principles of interleaving without explicitly doing so, or doing so under the banner of spacing or comparison. There is value, therefore, in considering the results across these sections side-by-side.

Given the small size of the evidence-base in this area, we have grouped all studies into a general interleaving strategy group, which (*ex-ante*) was judged to present sufficient evidence to examine the

effectiveness of the interleaving strategy. The discussion in this area looks more closely at the distinction between spacing, comparison, and interleaving, the practical issues around implementing interleaving, and its applicability across subject areas and contexts.

Main findings

Strategy 3: Interleaving

Concise definition

Interleaving involves switching between tasks or topics requiring different, but usually related, knowledge and skills. Like spacing, interleaving can take place within or across lessons.

Full definition and description

Interleaving involves switching between tasks or topics requiring different, but usually related, knowledge and skills. Like spacing, interleaving can take place within or across lessons. Tasks and topics selected for interleaving usually have common component ideas, solution steps, or other similarities. Interleaving is thought, therefore, to support discrimination between the tasks or topics or between the superficial and 'deep' aspects within each.

Selected examples

Examples of this strategy from our database include:

- In Rohrer et al. (2014), students received ten mathematic assignments across nine weeks with 12 problems in each. There were four kinds of problem, all involving algebraic strategies such as isolating terms, creating proportions, and finding possible values for terms. Problem types were either interleaved or blocked. The first four problems for each assignment were of one type. Then, for the blocked practice condition, the remaining eight were of the same type. For the interleaved condition, the remaining eight problems were distributed across the remaining assignments. (Also see Rohrer et al., 2019, 2020.)
- Rau, Aleven and Rummel (2013) examined interleaving of mathematics (a) problem types and (b) representations. The problems included diagrams representing fractions as either number lines, segmented circles, or sets. There were 12 task types all related to fractions; these included identifying fractions, making equivalent fractions, comparing fractions, and adding fractions. Students used a fractions 'intelligent' tutor programme for five hours spread across five to six days.
- In French, Rink and Werner (1990), high school students practiced basketball skills (forearm pass, set, or overhead serve) over nine class periods in three conditions that varied the blocking of the teacher presentation and the practice.

Evidence for this strategy

There were 12 high and medium priority studies of interleaving. Of these, six were graded as high relevance and quality. Full details of all medium and high studies are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed for this strategy are characterised as follows:

- **Pupil age and characteristics**. The age range of students was from third grade (age 8 to 9) to eighth grade (age 13 to 14). There was a roughly even split between studies of upper primary (age 8 to 11) and early secondary (age 12 to 14).
- **Location**. Eleven of the 12 studies in this area were conducted in the U.S. There was one study from Germany.
- **Learning areas**. Eleven of the 12 studies were of maths. There was a range of maths topics including fractions, algebra, subtraction, and geometry. The other study was in physical education (volleyball skill, in a study already reported in the spaced practice section).
- **Outcome measures**. All but one of the outcome measures were items developed by researchers aligned to the specific areas of instruction. In the case of computer programme-based interventions, the software also provided the outcome measure. There was one standardised measure used (for the study of volleyball). Some studies had conducted validity assessments of their outcome measures or based test items on common standardised tests.
- **Design and delivery**. Of the 12 studies, four were delivered via computer software, with students working independently. The majority (11) used either computer programmes or workbooks (with interleaved problems). About five included short, scripted instructional periods from teachers or an opportunity for teachers to provide feedback. Four were delivered by researchers and about five were delivered by teachers or with some teacher involvement (which included teachers facilitating and providing feedback on the computer and workbook-driven approaches). Overall, this is a highly scripted, researcher-controlled set of studies, driven mostly by workbooks or computer software.

High priority studies in this area

There were six studies in the interleaving category rated as having high strength and validity of evidence. We conducted in-depth analysis of these studies and have completed a full risk of bias assessment. We summarise these below and refer interested readers to the appendix for this section or the underlying paper for further details.

Booth et al. (2015). This study employed an experimental design to test the effect of the AlgebraByExample approach. The study included 380 eighth-grade students in 28 classes in five school districts in the U.S. The trial conditions were randomised at the class level. Treatment students received a workbook containing interleaved worked examples and self-explanation prompts. There was a mixture of correct and incorrect worked examples. Control students were given the same problems to solve. The content was taught by the regular maths teacher throughout. They assessed outcomes using assessments of conceptual and procedural knowledge (66 items, researcher-developed) and ten items from standardised algebra curriculum tests.

Key findings. Students receiving the AlgebraByExample intervention received higher post-test scores for standardised test items and conceptual knowledge. The effect of the intervention was especially strong for conceptual post-test scores for students with low prior knowledge. Treatment students outscored control students by 7% on the items from the state standardised test. For students in the lower half of the performance distribution this increased to 10%. Treatment group gains were also seen on the assessments of conceptual and procedural knowledge of 5% and 4%, respectively. The risk of bias assessment identified some concerns with the randomisation process, missing outcome data, and selection of the reported results. This was mostly an issue of report and protocol. The latter, for example, requires studies to have pre-determined statistical analysis plans. Overall, this was graded as ‘some concerns’.

Nemeth et al. (2019) investigated the flexible use of algorithmic and number strategies in elementary school maths (subtraction). Their focus was on whether an interleaved approach would improve student's flexibility and choices for strategies (for example, identifying 'shortcut' strategies and selecting the most suitable one for the task). The study was a 2 x 4 factorial design with two group conditions (interleaved versus blocked) and four time conditions for testing—before intervention, one day later, one week later, and five weeks later. The study involved 236 German third-grade (age 8 to 10) pupils from 12 classes in four elementary schools. Within each class, students were randomly assigned to either an interleaved or blocked condition. In the interleaved condition the students had to choose an appropriate strategy based on each task. Comparison processes were supported by prompting the students to compare the strategies (between-comparison) while the students of the blocked approach were encouraged to reflect the adaptivity of a specific strategy for specific subtraction tasks (within-comparison). Both groups were taught to use different number-based strategies (that is, shortcut strategies and decomposition strategies) and the standard written algorithm for solving three-digit subtraction problems spanning a teaching unit of 14 lessons. During the intervention, no regular mathematics lessons were held. The intervention was delivered by four trained staff members. The outcome measures were focused on strategy-use adaptability rather than overall performance. To assess the students' flexibility, the Flexibility and Strategy-Specific Adaptivity Test was used: their strategy use was coded by four trained coders independently guided by a standardized coding manual. This aspect of the study was borderline in terms of our eligibility criteria about learning outcomes. We judged that increasing student strategy selection is a realistic learning outcome, and discuss this study here, but provide this study in the main outcome table below for information only.

Key findings. This study suggests that (a) an interleaved approach extended by prompts is practicable and can be well integrated into regular elementary school classrooms. Moreover, that (b) it enhances the flexible and adaptive use of subtraction strategies among third graders compared to a blocked approach with prompts for within-comparisons. The risk of bias assessment identified some concerns with the randomisation process, missing outcome data, and selection of the reported results. This was mostly an issue of report and protocol. The latter, for example, requires studies to have pre-determined statistical analysis plans. Overall, this was graded as 'some concerns'.

Rau et al. (2013) studied the effect of interleaving multiple representations versus tasks types for fractions learning. This was an RCT with student-level assignment. A total of 158 fifth- and sixth-grade students (age 9 to 12) in 16 classes in three U.S. schools were involved in the study; the results are based on 101 students. The intervention was delivered using a web-based intelligent tutoring system. The researchers assigned students randomly to one of two conditions: the 'int-types' condition, where the *task types* were interleaved while the *graphical representations* were blocked, and the 'int-reps' condition, where the *graphical representations* were interleaved while the *task types* were blocked. Students in both conditions worked on the same 102 fraction tasks at their own pace with the help from the intelligent tutoring system. All learning tasks involved a single graphical representation. Each problem also involved the symbolic representation of fractions and a problem statement in text, but this was kept constant across conditions. Researcher designed tests of (a) representational and (b) operational knowledge. Each of the two test scales included both familiar and unfamiliar tasks. Two different equivalent versions were created with equal difficulty and students were randomly assigned to either one at pre-test or received the other at post-test.

Key findings. The results revealed that the int-types condition was significantly more effective (d = 0.33, 95 % CI: -0.05, 0.73) and more efficient (d = 0.37, 95 % CI: 0.02, 0.76) than the int-reps condition for representational knowledge, but no more effective or efficient than the int-reps

condition for operational knowledge. The risk of bias assessment identified some concerns with the randomisation process, missing outcome data, and selection of the reported results. This was mostly an issue of report and protocol. The latter, for example, requires studies to have pre-determined statistical analysis plans. Overall, this was graded as 'some concerns'.

Rohrer et al. (2014) tested the effect of interleaving mathematics tasks types. They randomised conditions at class level, balancing teachers to ensure that at least one of each participant teacher's classes was included. The study involved 140 seventh-grade pupils from one middle school in the U.S. There were three teachers and eight of their classes involved. Students learned to solve four kinds of problems drawn from their course: One group interleaved their practice of problems of type A and B and blocked their practice of problems of type C and D; the other group did the reverse. Across all assignments, the students saw 12 problems of each of the four kinds. Shortly before the scheduled date of each assignment, teachers received paper copies for their students and a slide presentation with solved examples and solutions to each problem. Researchers asked teachers to present the examples before distributing the assignment. A researcher-designed test was used as the outcome measure, with items aligning to the studied problems, although all of the test problems were novel. The test included three problems of each of the four kinds, and each of the four pages included a block of three problems of the same kind.

Key findings. In terms of results, a repeated measures comparison of the two halves of the test showed that interleaved practice was nearly twice as effective as blocked practice. The effect size was large, d = 1.05 (95 % CI: 0.80, 1.30). This benefit of interleaving was observed for each of the four kinds of problems. The risk of bias assessment identified some concerns with the randomisation process and selection of the reported results. This was mostly an issue of report and protocol. The latter, for example, requires studies to have pre-determined statistical analysis plans. Overall, this was graded as 'some concerns'.

Rohrer et al. (2015)—like the previous study—looked at interleaving in mathematics task types requiring strategy selection. Students in nine classes taught by three mathematics teachers were randomised into two cohorts with different testing schedules: one cohort was tested after one day, the other after a delay of 30 days (n = 63 for each group). Each class, therefore, included students at both test delays. Classes were randomised to study conditions. The study involved 126 seventh-grade students. A practice schedule was designed to produce a within-subject variable where students in one cohort received interleaved practice of graph problems and blocked practice of slope problems; the other cohort received the reverse. The study consisted of ten practice assignments, a review session, and a test. Each practice assignment consisted of 12 problems presented on two sides of a single sheet of paper. The ten assignments included 12 graph problems and 12 slope problems and the remaining problems were drawn from unrelated topics. Teachers presented the related topic tutorial immediately before giving the first two assignments, however, the scheduling of the remaining eight graph and eight slope problems varied. On the due date for each assignment, teachers presented the solution to every problem with the aid of a slide show created by the authors. As teachers presented the solutions, students were asked to correct their errors. There was a researcher-designed test aligned to the material. No problems had appeared in either a practice assignment or the review.

Key findings. Compared with blocked practice, interleaved practice produced higher scores on both the immediate and delayed tests: *d* = 0.42 (95 % CI: 0.07, 0.77) and 0.79 (95% CI: 0.43, 1.15), respectively. There were positive effects for both topics but only one was statistically significant. The risk of bias assessment identified some concerns with the randomisation process and selection of the reported results. This was mostly an issue of report and protocol. The latter, for example, requires studies to have pre-determined statistical analysis plans. Overall, this was graded as 'some concerns'.

Rohrer et al. (2019)—following on from their 2014 and 2015 studies and again looking at interleaving in maths requiring strategy selection—conducted a cluster randomised controlled trial assigned at class level. This was the largest study of the three involving 787 seventh-grade pupils in 54 classes from five schools in the U.S. The study included three parts: a practice phase with eight worksheets, a review worksheet, and a test. The entire procedure lasted about five months and the time course varied slightly across teachers. The worksheets included critical problems and filler problems: the critical problems were like the kinds of problems seen on the test; the filler problems were drawn from topics unrelated to the critical problems and were included partly to prevent students and teachers from inferring the difference between the two conditions. Students completed the worksheets during class under the supervision of their teachers. The teachers were able to provide one-on-one help to students while they worked on the problems. Teachers then presented solutions (after 30 minutes) and students had the opportunity to ask questions before correcting errors. The test included four graph problems (page one), four inequality problems (page two), four expression problems (page three), and four circle problems (page four). These were blocked so as to not advantage the interleaving group. None of the test problems had appeared previously in the study.

Key findings. One month after the original test, students took an unannounced test and the interleaved group outscored the blocked group, 61% versus 38% (*d* = 0.83; 95% CI: 0.68, 0.97). There was a positive effect for each of the 15 teachers (d = 0.23–1.48) and for each of the four kinds of problems. The risk of bias assessment did not identify any concerns with this study, which was judged to have a 'low' risk of bias across all categories.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section we focus on the study outcomes, summarised in Table B2.2. Studies identified as high relevance and quality have been marked with an asterisk.

Table B2.2: Interleaving (general)—summary of evidence

Study	Focus	Population	Finding
		High Priority Studies	
*Booth *et al.* (2015a)	Effect of AlgebraBy-Example assignments on algebra test scores	*N* = 380 8th grade 5 school districts, 28 classes US	***Positive*** • Students receiving AlgebraByExample intervention received higher post-test scores for standardised test items and conceptual knowledge (standardised effect in random effect model = 0.06, SE = 0.03, p < 0.05). • Effect of intervention was especially strong for conceptual post-test scores for students with low prior knowledge
*Nemeth *et al.* (2019)	Flexible use of algorithmic and number strategies in elementary school maths (subtraction)	236 German 3rd grade (age 8-10) pupils from 12 classes in 4 elementary schools.	*Performance not the focus. Study included for information on wider outcome.* • The results of this study suggest that an interleaved approach extended by prompts to compare (1) is practicable and can be well integrated into regular elementary school classrooms. Moreover, (2) it enhances the flexible and adaptive use of subtraction strategies among third graders compared to a blocked approach with prompts for within-comparisons.
*Rau *et al.* (2013)	The effect of interleaving multiple representations versus tasks types for fractions learning.	N = 158 5th and 6th grades students (age 9-12) (results based on N=101) 16 classes, 3 schools. US	***Positive for task type (strategy) interleaving, but not representations.*** • Int-types condition was significantly more effective (*d* = 0.33, 95 % CI = -0.05,0.73) and more efficient (*d* = 0.37, 95% CI = 0.02, 0.76) than the int-reps condition for representational knowledge. • Int-types condition was no more effective or more efficient than the int-reps condition for operational knowledge.

*Rohrer *et al*. (2014)	Interleaving in mathematics task types requiring strategy selection	N = 140 7th-grade pupils, 1 middle school, 3 teachers and 8 of their classes US	***Positive for task type (strategy) interleaving, but not representations.*** • A repeated measures comparison of the two halves of the test showed that interleaved practice was nearly twice as effective as blocked practice. The effect size was large (*d* = 1.05, 95 % CI = 0.80, 1.30). This benefit of interleaving was observed for each of the four kinds of problems.
*Rohrer *et al*. (2015)	Interleaving in mathematics task types requiring strategy selection	126 7th grade students 3 mathematics teachers and 9 of their classes participated. US	***Positive for task type (strategy) interleaving, but not representations.*** • Compared with blocked practice, interleaved practice produced higher scores on both the immediate and delayed tests (*d* = 0.42, 95 % CI = 0.07,0.77 and 0.79, 95 % CI = 0.43, 1.15, respectively). • There were positive effects for both topics, but only one was statistically significant.
*Rohrer *et al*. (2019)	Interleaving in mathematics task types requiring strategy selection	N = 787 7th-grade pupils 54 classes from 5 schools US	***Positive for task type (strategy) interleaving, but not representations.*** • One month later, students took an unannounced test, and the interleaved group outscored the blocked group, 61% versus 38% (*d* = 0.83, 95 % CI = 0.68, 0.97). • There was a positive effect for each of the 15 teachers (*d* = 0.23–1.48). • There was a positive interleaving effect for each of the four kinds of critical problems
Larger Studies (pupil n > 500) (Medium Priority)			
There were no larger studies at the medium priority level			
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
French *et al*. (1990)	Effect of spaced practice on volleyball skill	*N* = 139 1 high school, 4 classes, US	***Neutral*** • No difference on practice schedule between groups, for any volleyball skills
Patel *et al*. (2016)	Interleaving versus blocking fraction addition and multiplication practice	2 experiments N = 70 (two conditions) and N = 118 (3 conditions) 6th grade, US	***Mixed (may be more of a sequencing effect)*** • Across both experiments, blocked fraction addition-to-multiplication practice produced less learning than both interleaved practice and blocked fraction multiplication-to-addition practice. Differences between interleaved and blocked (multiplication-addition) conditions were small.
Rau *et al*. (2014)	Interleaved versus blocked sequences of multiple representation of fractions.	N = 474 (only N = 230 analysed) 4th- and 5th-grade students from 6 schools (31 classes) US	***Neutral*** • The results show that there was no significant main effect of practice schedules on any knowledge type (*d* = 0.37, 0.88, 0.09, 0.21), indicating that there was no global effect of practice schedules across immediate and delayed post-tests. • Some qualified support in post hoc analysis for marginal advantage of interleaving for consistency of learning across all areas and tests.
Todaro *et al*. (2019)	Contextual, concrete, or abstract example manipulations in interleaved vs. blocked sequences in maths	Three experiments N = 121, 34, 54 grade 5,4, 4 4, 1, 1 school(s) US	***Positive for task type (strategy) interleaving, but not contextual information*** • Experiments one and two found that interleaving math procedures is more important to learning than interleaving contextualized examples (*d* = 1.28). There were some negative effects for the latter (*d* = 0.40). Experiment three indicated that working memory predicted learning whereas presentation or example type did not. It is likely that decreased spacing between interleaved math procedures attenuated the interleaved effect in Experiment 3.

Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Todaro *et al.* (2017)	Interleaving geometry problems and contexts	N = 37 (N = 33 for analysis) 4th grade US	***Positive for task type (strategy) on procedural knowledge.*** • When math skill was interleaved (i.e., interleaved and hyper-interleaved groups), procedural performance on post-test was significantly better than blocked (i.e., context interleaved group) (*d* = 1.24,1.37, 95 % CI = 0.31,2.14) • A significant effect was not found for declarative knowledge.
Wagner *et al.* (2019)	Effect of interleaved practice vs. repetitive practice and incremental rehearsal when learning single digit addition and multiplication	N = 74, 3rd and 4th grade students 1 School US	***Positive*** • Results indicated very few differences between practice conditions regarding acquisition accuracy, increased accuracy during retention trials for interleaved and incremental rehearsal practice, and higher learning efficiency for interleaved practice when compared to incremental rehearsal.

* High priority study identified for in-depth analysis.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in the table below.

Table B2.3: Interleaving—quality of evidence assessment (based on the GRADE approach)

Strategy	Interleaving (general)
Number of Studies	There are 12 studies in this area, of which six were rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	All studies are randomised experiments.
Risk of bias	Our risk of bias assessments on the high-quality papers identified some concerns with the randomisation of five papers, missing outcomes of three, and selection (or non-pre-specification) of five. One paper had low risk of bias in all areas. We judge there to be at least three strong studies in this area.
Inconsistency	**Result consistency**. The results are moderately consistently positive. There were eight positive results and three mixed/neutral results.
Indirectness	**Practice heterogeneity**. With only a couple of exceptions, the results have focused specifically on interleaving of maths tasks requiring different solution strategies. The specific mathematics content varied, as well as details of the procedure. These differences are more granular than this review is designed to explore so these studies are considered highly homogenous. **Population, measure, and outcome heterogeneity**. The studies spanned a range of ages from 8 to 14, and the vast majority were in maths. Generalisations beyond these ages and maths are not possible based on these results. **Outcome measures**. Outcomes were mostly researcher-designed tests aligned to the specific content targeted in the instruction. **Design and delivery**. Overall, this is a highly-scripted, researcher-controlled set of studies, driven mostly by workbooks or computer software.

Imprecision	**Group sizes**. The sample of studies included three moderate to large studies (n > 350), seven moderate-small studies (349 > n > 100) and two smaller studies. The number of studies is still relatively low for judgements of this type. There are several medium to large studies that suggest moderate to large effect sizes: - *Rau et al. (2013): *d* = 0.33 (95% CI: -0.05, 0.73); - *Rohrer et al. (2014): *d* = 1.05 (95% CI: 0.80, 1.30); - *Rohrer et al. (2015): delayed tests, *d* = 0.79 (95% CI: 0.43, 1.15); - *Rohrer et al. (2019): *d* = 0.83 (95 % CI: 0.68, 0.97); and - Rau et al. (2014): *d* = 0.37, 0.88, 0.09, and 0.21.
Publication bias	There are two smaller studies, both positive and one with large effect sizes (the other not reported). While this provides a slight suggestion of publication bias, the rest of the results provide no further indication that it might be present.
Other considerations	Of the six high priority studies, three are from the same author (Rohrer). Rau provides two studies and one positive result, one negative. Todero provides two studies with two positive results. This overlap in authors, as well as the focus on mathematics learning, reduces confidence that these results will apply in different contexts.
Overall confidence	**Low (++)** Our confidence in the effect estimate is limited: the true effect may be substantially different from our estimate.
Confidence reasons	This low confidence is based on the following issues: - Our intention was to evaluate interleaving across the curriculum; 11 of the 12 studies in this area, however, were in maths. - The highest precision studies, based on full reporting of effect sizes and confidence intervals, and of medium to large scale (including four high priority studies), while providing promising results were conducted by two authors. - The ecological validity evidence pointed to this being a highly scripted, researcher-controlled set of studies, driven mostly by workbooks or computer software. - We cannot be confident that the reported result applied outside the specific circumstances and learning objectives of the key studies on which the overall finding rests.

Summary of findings for this strategy

Main finding. Overall, the evidence provides support for interleaving as an effective approach for upper primary and lower secondary mathematics. There is insufficient evidence beyond this specific subject and age range.

Estimated impact. The highest precision studies suggest that effects for the specific learning objectives in upper primary and lower secondary mathematics may be moderate to high, with effect size estimates, d, between 0.33 and 1.05.

Confidence in impact estimate. We have rated our confidence in this result as low. Even within the confines of maths and the age range in question, there are limitation of ecological validity and breadth in the evidence, especially in relation to authorship, subjects, age, and learning objectives.

Heterogeneity. The evidence was insufficient—in both variation and scale—to examine variation in the effects by subgroups or factors.

Other points. This result should be examined alongside Strategy 9—schema or concept comparison and cognitive conflict—due to the similarities in theoretical rationale for the strategies.

Interleaving—overall evidence summary and conclusions

Summary of results

In this section, we have reviewed 12 studies focused on interleaving We grouped these into a general interleaving area for which we potentially had sufficient evidence to assess effectiveness. Our results for these are summarised in Table B2.4.

Table B2.4: Interleaving—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Interleaving	Twelve, of which six were graded as high priority.[1]	For specific applications of interleaving in maths (relating to tasks involving strategy selection) the overall evidence suggests moderate to large effect sizes.	The studies spanned a range of ages from 8 to 14 and the vast majority were in maths. Generalisations beyond these ages and maths is not possible based on these results.	Low (++)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Conclusions about strategies in this area

Interleaving

Our headline conclusions in this area are:

- The most notable aspect of this evidence-base is that 11 out of 12 studies were in maths. Moreover, even within these, the focus was on interleaving mathematic tasks that required learners to select a solution strategy before implementing it. This was true for five of the six high priority studies. The other study (Nemeth et al., 2019), while it did not test overall performance differences, found that interleaving improved the flexibility and suitability of students' strategies.
- The evidence supports the overall theory of how interleaved *maths* tasks may promote learning: with variation and the need to actively select strategies, students become more familiar with the differences between strategies, more able to discern between them, more confident in carrying them out, and more discerning and flexible at selecting them.
- For this specific application of interleaving, the overall evidence suggests moderate to large effect sizes.

A question not addressed in this data, to which we return in the discussion and questions section, is whether interleaving is likely to have value across other subjects and applications. We review literature recommending the application of interleaving across the curriculum and discuss the theory and practice in this area. Finally, we return to the opening comments in this section about the connection between interleaving and spaced practice in terms of timing and comparison (a strategy in the Working with Schemas section). Assessing evidence in these three sections side-by-side is valuable for increasing understanding of this area.

Evidence-informed discussion and questions

About this section

Across our study database, the practice review literature, and our exploration of teacher perspectives interleaving has been a concept that has arisen less often. The general grouping of interleaving strategies and the narrowness of subject areas represented in our main review, above, reflects this. This discussion section, therefore, necessarily draws on sources from the practice review and teacher perspectives from primary data to briefly outline some of the questions and challenges in this area that future research and practice might explore.

Principles and moderating factors

How is interleaving thought to work? How does interleaving differ from spacing? Is there a 'value-added' of interleaving as well as spacing learning?

Here, we briefly relay explanations for how interleaving works from the practice review literature. We draw heavily on Dunlosky et al. (2013) and Weinstein et al. (2018) whose accounts were more detailed than elsewhere. They primarily summarise findings from the cognitive psychology literature, which includes a mixture of laboratory studies and studies in more naturalistic settings.

Dunlosky et al. (2013, p.41) briefly consider why interleaving can be beneficial. They offer two explanations: first, that 'interleaved practice promotes organizational processing and item-specific processing because it allows students to more readily compare different kinds of problems'; second,

that interleaving, by nature, tends to produce a form of spaced retrieval practice. They also highlight research suggesting that interleaving can be beneficial over and above any spacing effect (Kang and Pashler, 2012; Mitchell, Nash and Hall, 2008), suggesting that both explanations are likely to be at play. The theme of comparison is common in practice-facing sources we consulted. The idea behind interleaving is described as supporting students to 'compare', 'contrast', 'differentiate', 'synthesise', create 'connections' between, identify 'variation' in, and 'discriminate' related concepts and strategies (Dunlosky et al., 2013; Weinstein et al., 2018; Agarwal and Bain, 2019; Brown, Roediger and McDaniel, 2014). Agarwal and Bain (2019, p.112) claim that 'the key to interleaving is discrimination', adding that the more similar items are, the harder discrimination becomes, and thereby interleaving similar but different items can be a 'desirable difficulty'. This general idea is evident in our main evidence review—and especially in relation to the ability of students to discriminate and select between different maths strategies. We note that there is an area of the education research literature in which very similar ideas are discussed under the term 'variation theory' (Lo, 2012). Beyond this central idea of comparison, the practice and the prevailing ideas about interleaving become less unanimous. Below we explore some of the variation and teacher perspectives on implementation.

Variation in the practice or teaching and learning context

What should be interleaved with what (and why)? Should teachers interleave topics, activities, solution strategies, retrieval practice items, or something else?

One key area of questions, both for the review team when reading the literature and posed by teachers in our interviews and surveys, relates to what exactly teachers should interleave and the relative merits of doing so. In our main evidence, the focus was almost entirely on the interleaving of maths strategies in a way that forced learners to select and use mathematic strategies. This would require students to discriminate between different problems and problem-solving strategies, providing a plausible mechanism for this effect. With the broadly positive results, there is therefore both evidential and theoretical support for interleaving of mathematic problems requiring different strategies (see Rohrer et al., 2014; and Rohrer et al., n.d.).[6]

We did not locate evidence as to whether the benefits of interleaving extend beyond strategy choices in maths. Weinstein et al. (2018, p.7) ask this question and suggest that 'the answer appears to be yes'. Studies such as Kang and Pashler (2012), a highly-cited study within the interleaving literature, have found benefits for students learning different artists' styles, suggesting that there may be wider benefits.

What is particular about maths and how might this affect interleaving in other subjects? How sensible is it to transfer this to all subjects?

Many teachers provided perspectives on the application of interleaving, often approaching the question in terms of which subjects or topics were more or less suited to it. Topics described as suitable included grammar in the language curriculum, physics equation practice, preparation for MFL oral exams (where 'students have to be ready to talk at length about a large number of topics'), maths teaching, and primary languages. One teacher thought that interleaving was effective for developing skills as well as knowledge for all year groups and that switching between skills helps both learning and engagement. Others gave more general descriptions such as interleaving being more valuable

[6] A guide, Interleaved Mathematics Practice, is available at the following location: http://uweb.cas.usf.edu/~drohrer/pdfs/Interleaved_Mathematics_Practice_Guide.pdf

'where you need to keep all the plates spinning at once', or—in the negative—that one must 'be careful about interleaving as some topics do not lend themselves to it very well'. Overall, the teacher perspectives and evidence we have do not provide a clear picture of what (if anything) should be interleaved beyond maths strategies. Practical advice certainly suggests that various things can be interleaved (for example, concepts, topics, categories, and strategies) but the evidence is at present quite thin in terms of what is more effective or the principles of applying interleaving in different areas. Future research that examines the benefits of interleaving in non-maths subjects in real classrooms would be highly valuable to the field.

Implementation

Interleaving strategies and their implementation

How are teachers applying interleaving (based on our interviews and survey responses)?

Teachers reported a wide range of activities and strategies as forming the basis of interleaved practice. Many of these related to the interleaved start-of-lesson or end-of-lesson activities, such as the following.

Lesson starter and closing activities

- 'I have created a bank of "exit questions" which meet our spec. Staff are to "cold call" these and pick the topics which interleave with the topic taught in the lesson.'
- 'I use interleaving in my starters and want to introduce it more in my independent practice.'
- '[I use interleaving in] starters and ends of lesson and with home learning.'
- 'We commonly use interleaving in practise but also as part of retrieval starters—again interleaving concepts that have been taught in previous units/weeks. Starters routinely return to previously taught work on a cycle to maintain rapid recall. This gives an opportunity to reteach specific concepts if children are unfamiliar.'
- 'I see each class twice per week. First lesson starter is ten questions. These are retrieval, interleaved, and space practice.'
- '[It involves] having old topics as starters and using curriculum links as an opportunity to review past topics.'
- 'We call them "wake up shake ups"—they have repeated style questions on them over a few weeks and then we move onto a different aspect of learning: time, angles, word meaning.'

There was some discussion around trying to build interleaving into the wider curriculum planning and schemes of work. Again, these responses were very similar to those reporting in relation to spacing. Indeed, retrieval practice, spaced practice, and interleaved practice were very often mentioned collectively.

Challenges of implementing interleaving

One thing that was quite striking about our bank of comments about interleaving relative to other areas was the proportion of negative comments relating to the difficulties of understanding and implementing it as a strategy. This is reflected in our practice review literature sources, with comments such as the following:

> *'On the negative side, the literature on interleaved practice is currently small, but it contains enough null effects to raise concern. Although the null effects may*

> *indicate that the technique does not consistently work well, they may instead reflect that we do not fully understand the mechanisms underlying the effects of interleaving and therefore do not always use it appropriately.'*
>
> *(Dunlosky* et al., *2013, p.44)*

Examples of comments from teacher interviews and questionnaire respondents related to interleaving being hard for teachers to understand include:

- 'We talked a lot early on about interleaving, but it is a bit hit and miss how that is interpreted in the school. It gets a bit jumbled up with spaced practice and retrieval, but not deliberately practiced as well as the other two' (Interviewee 8).
- 'Interleaving is something that they [ITT trainees] always struggle with; the idea that they are trying to develop the understanding of the sequence and having to think about how they're revisiting concepts and how that affects long term retention. And the depth of understanding you need to have to grasp how interleaving in theory can work is a lot more challenging than retrieval practice.'
- 'In language teaching I try to use all [strategies]. I have misunderstood interleaving as interleaving within a topic rather than within a lesson … I was trained to vary the activities within a lesson—I didn't realise till just now that that was interleaving!'
- 'I find interleaving hard. [I] tend to concentrate on one main concept in a lesson and trying to get the explanation really clear rather than swapping around.'
- 'Possibly because at A-level Literature we are constantly developing the same complex skills and it feels disruptive to chop and change from one text to another. I'm experimenting with it by doing half the teaching on each text in year one of A-level and then returning in year two, interleaving with unseen texts.'
- 'Interleaving is difficult to implement: it overlaps with spaced practice and it's not easy to choose which subjects to interleave.'
- 'I just can't get my head around how to make it work. It seems like it would be confusing for the students jumping around (although arguably teaching biology/chemistry/physiology on rotation may inadvertently be providing this?)'
- 'I don't think anyone I work with (including myself) knows how to use it properly.'
- '[It is] difficult to know how much time to spend before switching idea/topic, i.e., the granularity of the unit'
- 'I am not yet very successful with interleaving but that might be a problem with me rather than the approach. The children need time to embed and link and too much 'swapping' has meant that some learning is lost; as I say, I think I haven't been successful yet! But I will keep returning to this.'
- 'I am struggling to do this so it is effective. I have tried to have interleaved homework, but I know that this is just a bad attempt at interleaving. I am not entirely sure how this one is effective.'
- 'I find interleaving can sometimes be contradictory, with overloading. My interleaving works best to connect ideas together so although it might be content from one topic area, it links well to another. Still developing this.'
- 'Interleaving is difficult as there is a balance to be struck between giving children sufficient time to acquire and embed a skill before interleaving with other skills.'
- 'Interleaving overcomplicates lesson design and primary staff are not sufficiently subject specialists to make effective—what is difficult to implement will not be successful.'

The area of interleaving received a far greater number and proportion of negative comments than the other topic areas in our review. There were also a range of comments relating to the students being confused by interleaving.

> *'Students really struggle with interleaving because, say if I'm teaching two literary texts like* Macbeth *and* An Inspector Calls*, interleaving would kind of suggests that if I get him to start making comparisons between themes that would bring him to a deeper understanding but students don't want to do that. They want to compartmentalise and are like, "Miss, why do we have to write about these at the same time? Why are we comparing them?" and so sometimes I have to try and persuade them that active comparing them will give them a deeper understanding, give them fresh insights, even though in the exam they would be dealt with separately.'*
>
> *(Interviewee 4)*

Similarly, many teachers talked about potentially confusing students, especially students with lower attainment (also see Dunlosky, 2013, p.42). A few teachers thought that interleaving was more appropriate for higher-attaining students who can 'cope with the synoptic approach'. Related, Weinstein et al. (2018) suggest that teachers should proceed with caution when promoting interleaved independent study.

The final group of teacher comments on implementation (again, mostly negative) related to the difficulties of planning and timetabling for interleaving. Similar issues to those discussed for spaced practice were apparent: the curriculum being too content-heavy to allow for repetition of non-core topics, the difficulties of interleaving in the context of rigid exam syllabi and school timetables, and the challenges for workload around planning and lesson preparation.

Final thoughts on this strategy area

In our systematic review of classroom trials, we concluded that interleaving *maths* tasks may promote learning. As discussed further in this section, the principles of discrimination, comparison, and connection seem to be at play. Encouraging students to select strategies actively in maths fits this explanation. Moreover, several studies found substantial effect sizes, including studies we rated a having a low risk of bias (Rohrer et al., 2019, with an effect size of d = 0.83). While the limited evidence led us to rate this as a 'low confidence' judgement, the evidence is promising for this particular subject and application. Beyond this, interleaving and its benefits (over and above spaced practice) appear less certain, and we—as per Dunlosky et al. (2013, p.44)—feel that this is an area where our understanding of the theory and practice is currently relatively limited.

B3. Retrieval practice

Overview of area

Definitions

Retrieval practice 'refers to the act of recalling learned information from memory (with no or little support)' (Jones, 2019). Principles of learning from cognitive science suggest that learners actively generating responses from memory and quickly receiving feedback will be an effective learning approach. A common way of achieving this in a classroom is through low-stakes quizzes, questions, and tests. As Dehaene (2020) explains:

> *'Regular testing maximises long-term learning. The mere act of putting your memory to the test makes it stronger. It is a direct reflection of the principles of active engagement and error feedback. Taking a test forces you to face reality head-on, to strengthen what you know, and to realize what you don't know.'*
>
> *(Dehaene, 2020, p.214)*

The idea that taking a test might be a good strategy *for* learning as well as an assessment *of* learning is not immediately obvious or intuitive. One might reason that either the student knows the answer or they do not, and testing assesses this without affecting their knowledge base. However, cognitive science highlights that memory has a 'strength' and, over time, a memory's strength diminishes.[7] Recall—or retrieval—of information strengthens memory. Seminal studies have found that testing can be more effective than restudying the same material (Roediger and Karpicke, 2006). Restudy makes the material feel more familiar and students perceive it to have resulted in learning. On the contrary, Roediger and Karpicke (2006) found testing to be more effective because it retrieves the knowledge from long-term memory rather than simply re-presenting it to the working memory. Thus, retrieval practice—like spacing and interleaving—can be viewed as a 'desirable difficulty': while students find retrieving information more difficult than simply restudying that information, it promotes long-term retention.

In this section we take restudy or an instructional recap (both a form of re-presentation of the material) as the alternative condition or strategy against which retrieval practice should be evaluated. However, not all studies have enabled this comparison. We have grouped all studies in a general strategy group, testing this general principle behind retrieval practice, sometimes referred to as 'the testing effect'. We note that the form that the test takes varies considerably. Commentators and translators of cognitive science for practitioners such as Sherrington and Caviglioli (2020) provide examples including multiple-choice questions, short-answer fact questions, short problem solving (for example, solving simple sums), true/false questions, error spotting, labelling diagrams, image recognition, recitation of quotes or definitions, and list creation (also see Jones, 2019). Moreover, there appear to be appreciable variations in practice relating to the provision of accompanying feedback, differentiation, or targeting of test items, and variation of learning content to promote transfer. We will discuss these in the discussion and questions section following the main results.

[7] See for a concise introduction: Yan, V., (n.d.) 'Retrieval Strength vs. Storage Strength', Learning Scientists Blog, Guest post. Available: https://www.learningscientists.org/blog/2016/5/10-1

Overview of the evidence-base

Table B3.1: Retrieval practice—overview of study priority ratings

Priority level	Overall rating	Ecological validity	Relevance and definition for focus CS practices	Added value to evidence-base
High	4	2	28	19
Medium	35	32	29	40
Low	26	31	8	6

The review database contained 65 studies in the retrieval practice category. Of these, 39 were graded as being of sufficient ecological validity, relevance, and value for inclusion within this analysis of the evidence (high and medium). There were (only) four studies that scored highly across these criteria and were identified as *potentially* providing strong evidence in this area (high).

The overview of priority assessments in Table B3.1 shows that ecological validity of the studies in the section tended to be medium to low. As we discuss further below, many studies were highly contrived researcher set pieces, with narrow learning objectives, designed to provide a strong (that is, internally valid) experimental test of retrieval practice rather than an ecologically valid one. Relevance and adherence to the definition of retrieval practice as a concept were high to medium, with many studies offering relevant strategy tests in a classroom context.

This section has grouped all studies in a general strategy group, all testing the general principle behind retrieval practice, sometimes referred to as 'the testing effect'. We take restudy or an instructional recap (both a form of re-presentation of the material) as the alternative condition or strategy against which retrieval practice should be evaluated.

Wider evidence in this area looks at questions such as the role of feedback in retrieval, the impact on cognitive load, the use of hints, cues or prompts, variations in test formats; the personalisation or targeting of practice, the transfer of learning via retrieval, and the influence of prior learning on retrieval success and value.

Main findings

Strategy 4: Retrieval practice (compared to restudy)

Concise definition

Retrieval practice refers to any activity that requires students to recall information from memory rather than representing or restudying the information.

Full definition and description

Retrieval practice refers to any activity that requires students to recall information from memory rather than recapping, revising, or restudying the information. This can include partial recall of information supported by hints, cues, scaffolds, or other contextual information. A common way of achieving this in a classroom is through low-stakes quizzes, questions, and tests.

Selected examples

Examples of this strategy from our database include:

- Agarwal (2019, experiment three) used twelve, four-alternative, multiple-choice fact questions and twelve, four-alternative, multiple-choice, higher-order questions for each of two textbook chapters ('Russian Revolution' and 'World War 2') in a social studies textbook. 'Each chapter unit lasted approximately one week. For each unit, students read a chapter from their social studies textbook, listened to seven or eight lessons, participated in quizzes (the experimental manipulation), and completed standard assignments developed by the teacher' (p.200).
- In Damhuis, Segers and Verhoeven (2015), retrieval practice consisted of repeated testing of vocabulary recall after a storybook reading (20 minutes). This was compared to a repeated storybook reading condition and a retrieval practice (the child had to choose the correct picture from a choice of four to represent a given word) with feedback condition (providing the correct answer for the picture task irrespective of whether the child was correct or incorrect).
- There was a large range of retrieval tasks across the studies including gap-fill tasks, matching tasks, multiple choice tasks, factual questions and true-false questions, and writing notes from memory (turning over a factsheet and recall on the back). These were mostly short-answer questions in the form of quizzes.

Evidence for this approach

There were 21 studies that examined the 'testing effect' of retrieval practice. Of these, three were graded as high relevance and quality (Nb. one of the four high quality studies in the overall section was not part of a strategy group with a sufficient weight of evidence to assess the strategy). Full details of all medium and high studies are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed for the retrieval practice strategy are characterised as follows:

- **Pupil age and characteristics**. The age range of students spanned from the early years (age 4–5) to age 16–17. There was a good spread of studies across this range, although most studies (15) were of students between the ages of 8 and 14, with two below this and four above.
- **Location**. There were a range of countries represented in the data, though the majority were north-western Europe or the U.S. The figures were as follows: the Netherlands, five; Germany, two; Australia, one; Brazil, one; Sweden, one; the U.S., nine; and the U.K., two.
- **Learning areas**. A good range of learning areas were examined within the studies. There were six studies of vocabulary learning, two in the early years; seven studies of history, geography, social studies, psychology, or thematic topics containing these; four science-focused studies; one maths; one literacy (text writing); and two with various subjects tested. The learning outcomes tended to be factual recall or vocabulary learning. However, there were a small number of examples of learning with higher 'element interactivity', where elements needed to be connected as well as recalled (for example, located on a mind-map).
- **Outcome measures**. The vast majority of studies (18 of 21) used researcher-designed tests directly aligned to the targeted learning content. The targeted learning content was, in most cases, based explicitly on regular curriculum content. There was some variation in the timing of these tests, with many using both immediate and delayed tests. There were two examples of the regular end-

of-unit assessments being used, and one use of a standardised (psychometric) test, the Maze Reading Comprehension Test, alongside a researcher-developed test of targeted word learning.

- **Design and delivery**. One ecological validity issue in this area is that of the 21 studies, only one was reported as being delivered by regular class teachers. Researchers or research assistants delivered the other 20, sometimes in the classroom, sometimes in small groups in the school but out of the regular class. Of these, several used technology-based quizzes and scripts to standardise the delivery of the intervention.

High priority studies in this area

Three studies investigating retrieval practice were rated as having high strength and validity of evidence. We conducted in-depth analysis of these studies and have completed a full risk of bias assessment, summarised in the appendix.

Agarwal (2019; experiment three only). The third experiment in this study employed a 3 x 2 randomised design and was conducted with a sample of 142 sixth-grade students from one school and six classes in the U.S. The study examined the effect of retrieval practice and question type on higher-order learning in history. There were three retrieval practice conditions (higher order quizzes, mixed quizzes, and non-quizzed). Unfortunately, there was no restudy condition so this study has been included in this section for comparison purposes only. There were two test conditions (fact test and higher-order test). The quizzing items were focused on textbook chapters chosen by classroom teachers. During the experiment, the students were quizzed in-class using a clicker-response system facilitated by the researchers. The study lasted for approximately two weeks. The outcome measure consisted of 24 multiple-choice questions based on quizzed content. This was researcher-developed, but with content based on the social science curriculum.

Key findings. Overall, the mixed retrieval practice condition produced the greatest level of performance on both fact (d = 1.44–1.55) and higher order final tests (d = 0.34–1.37). The risk of bias assessment identified some concerns with the randomisation process and selection of the reported results. This was mostly an issue of report and protocol. The latter, for example, requires studies to have pre-determined statistical analysis plans. Overall, this was graded as 'some concerns'.

Churches et al. **(2020)** was a meta-analysis of 34 small teacher-led RCTs of which 21 were focused on retrieval practice. For studies across all cognitive science areas, teachers were provided with an RCT design day and pre-reading material about RCT design and cognitive science concepts. Teachers then designed and led their own RCTs. Curriculum subjects included mathematics (times tables, problem solving), English (vocabulary, spelling), science, history, and geography. Trial length varied from a single lesson to 42 days. For retrieval practice specifically, there were 21 studies; these were not reported in detail individually but details of their topic area, n, effect size, and an analysis of their robustness are provided. We reproduce an overview of the individual studies in Table B3.2.

Table B3.2: Retrieval practice trials in the meta-analysis of Churches et al. (2020)[8]

Author	Year group	Subject	n	Effect size (d)	Jadad score for robustness (0–5)
Morris (2018)	4	Maths (times tables)	60	1.15***	3
Dunford and Rhoades (2018)	3	Maths (times tables)	25	0.87*	2
Dunford and Rhoades (2018)	2	Maths (times tables)	28	0.75*	2
Siddle (2018)	5	English (vocab)	37	0.65**	3
Siddle (2018)	2	English (vocab)	26	0.59*	3
Siddle (2018)	5	English (vocab)	36	0.58*	3
Pemberton (2018)	2	Maths (times tables)	24	0.58*	3
Elliot and Wyatt (2018)	4	Maths (times tables)	50	0.45*	2
Quinn and Lamb (2018)	8	English (vocab)	286	0.37**	3
Maberly (2018)	9	Science knowledge	92	0.32	2
Siddle (2018)	EYFS	English (vocab)	63	0.32	3
Greenfield, Noden and Siddle (2018)	4	Maths (times tables)	223	0.23	3
Siddle (2018)	EYFS	English (vocab)	41	0.16	3
Siddle (2018)	2	English (vocab)	65	0.10	3
Siddle (2018)	3	English (vocab)	44	0.04	3
Makarova (2018)	10	Science knowledge	110	0.01	2
Baker and Hindley (2018b)	4	Maths (times tables)	91	-0.06	3
Siddle (2018)	3	English (vocab)	44	-0.10	3
Baker and Hindley (2018b)	5	Maths (times tables)	108	-0.18	3
Morris (2018)	4	Maths (times tables)	60	-0.30	3
Baker and Hindley (2018a)	4	English (spelling)	172	-0.82*	3

Key findings. Overall, these studies suggest a positive impact of retrieval practice for vocabulary, times tables, knowledge, and scientific knowledge. All are for primary age children. There is variation in the sample size, including two studies with n > 200 and three between 100 and 200. There are no details about why the effect sizes might vary given the ostensibly highly similar conditions. The risk of bias assessment for this study raised some concerns with the randomisation process, outcome measurement, and deviations from the intended intervention. Overall, this study was rated as having 'some concerns', which we interpret as presenting indicative evidence. We note that all studies are highly ecologically valid, by design, and the varied authorship and contexts support generalisation.

Roediger et al. (2011). This study tested the effect of quizzing on social study test scores using a within-subjects experimental design. There were three experiments: in experiments one and two, the study included 142 sixth-grade students from one middle school in the U.S. In experiment three, 132 sixth-grade students from the same school took part. In experiments one and two, performance on quizzed items was compared to that on items that were presented twice (experiment two) or items that were not presented on the initial quizzes (experiments one and two). In experiment three, students were given one multiple-choice quiz in class and all were encouraged to quiz themselves outside of class using a web-based system. Students studied material used as part of a regular social studies course on cultures in all experiments. End-of-semester exams were used as outcome measures.

[8] All references are provided in the appendices

Key findings. In terms of results, overall, students performed better on quizzed than non-quizzed items. In experiments one and two, students' performance on both chapter exams and semester exams improved following quizzing relative to either not being quizzed or relative to the twice-presented items. For experiment three, the quizzing of material produced a positive effect on chapter and semester exams relative to control conditions. The risk of bias assessment identified some concerns with the selection of the reported results. This was mostly an issue of report and protocol. The latter, for example, requires studies to have pre-determined statistical analysis plans. Overall, this was graded as 'some concerns'.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section we focus on the study outcomes, summarised in Table B3.3. Studies identified as high relevance and quality have been marked with an asterisk.

Table B3.3: Retrieval practice (compared to restudy)—summary of evidence

Study	Focus	Population	Finding
High Priority Studies			
*Agarwal (2019): Expt.3 only	Effect of retrieval practice and question type on higher-order learning in history	*N* = 142 6th grade 1 middle school, 6 classes US	***No restudy condition. Positive compared to non-quizzing.*** • Overall, mixed retrieval practice produced the greatest level of performance on both fact and higher order final tests: • Final fact test: mixed quiz condition resulted in far greater performance (91%) than the higher order quiz and non-quizzed conditions (64% each) (*d* = 1.44, t(47) = 12.24, p < .001 and *d* = 1.55, t(47) = 13.63, p < .001, respectively). • For the final higher order test, the mixed quiz condition again resulted in the greatest performance (82%) compared with the higher order (75%) and non-quizzed (56%) conditions (*d* = 0.34, t(87) = 2.27, p = .078 (p =.026 without Bonferroni correction), and *d* = 1.37, t(87) = 12.24, p < .001, , respectively).
*Churches *et al.* (2020)^	Teachers designed and led RCTs utilizing retrieval practice	*N* = 2,157 Early Years to Year 6 31 schools (34 teachers) UK	***Positive*** Retrieval-specific studies: • 16/21 retrieval practice studied showed positive effect size- 9 of these had statistically significant results. • 5 showed negative effect (only 1 of these was significant) (Y4 English Spelling) • Retrieval practice related protocols yielded a positive overall pooled effect (r = 0.14, 95% CI [0.06, 0.23] [d = 0.28], p = .001).
*Roediger *et al.* (2011)	Effect of quizzing on social study test scores	*2 experiments* *N* = 142 6th grade 1 middle school US Same participants used in both	***Positive (against twice-presented items)*** • Overall: students did better on quizzed than non-quizzed items • Expts. 1 and 2: Students' performance on both chapter exams and semester exams improved following quizzing relative to either not being quizzed or relative to the twice-presented items • Expt.3: quizzing of material produced a positive effect on chapter and semester exams • Pairwise comparisons confirmed a significant testing effect (tested greater than non-tested), (t(62) 7.60, d = .98), as well as a significant testing benefit relative to read items, (t(62) 6.61, *d* = .83)
Larger Studies (pupil n > 500) (Medium Priority)			
There were no larger studies at the medium priority level			
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
Damhuis *et al.* (2015)	Effects of repeated storybook reading versus testing on	*N* = 140 Ages 4-5 years 6 elementary schools, 11 classes	***Negative*** • Repeated storybook reading, which may be regarded as a restudy condition, was as effective as testing for stimulating the breadth of vocabulary knowledge but found to be even better than repeated testing for stimulating deeper vocabulary knowledge. (d = 0.65, 95 % CI = 0.15, 1.15)

	vocabulary learning	The Netherlands	• For the breadth and depth of vocabulary, feedback enhanced the positive effects of testing, but not until feedback was given at least two times. • Testing with feedback was not superior over listening to repeated storybook readings. A nonsignificant trend between these conditions might suggest that repeated storybook readings lead to greater depth of word knowledge than repeated testing.
Goossens *et al*. (2014b)	Effect of retrieval practice on vocabulary learning	Two experiments *N* = 147, 122 Aged 7-10 years 2 primary schools, 6 classes The Netherlands	***Positive for fill-in-the-blank test*** ***Neutral for multiple-choice test*** • Expt. 1: No significant difference in test scores between conditions • Expt.2: on recall, children in the retrieval practice condition outperformed the children in the pure restudy condition, (*d* = 0.88), and the elaborative restudy condition (*d* = 0.82). No effect of condition on recognition
Goossens *et al*. (2016)^	Effect of retrieval and spaced practice on vocabulary learning	*N* = 129 2nd, 3rd, 4th and 6th grade The Netherlands	***Neutral*** • No significant benefit of retrieval practice or spaced practice on either the cued-recall or multiple-choice tests ***Negative*** • Some significant effects in unexpected direction: benefits of restudy in Grade 3, and short-lag spacing in Grades 2 and 4
Hanham *et al*. (2017)	Effect of testing and element interactivity on learning to write types of text	*6 experiments* *N* = 48, 43, 19, 18, 18, 29 Age 8-12 years 1-2 schools, 1-2 classes Australia	***Positive*** • The testing effect on immediate tests was larger and more likely using lower element interactivity materials (i.e., lower cognitive load) (*d* = 1.37, 95 % CI = 0.04,1.65) ***Negative (but see discussion)*** • A reverse testing effect was likely on immediate tests tapping higher element interactivity material but possibly eliminated by using a delayed test (*d* = 1.03, 95 % CI = 0.39, 1.66)
Jagerskog *et al*. (2019)^	Effect of retrieval practice versus multimedia learning on psychology recall	*N* = 133 Aged 16-17 years 3 high schools, 5 classes Sweden	***Neutral*** • Main effect of presentation format: visuo-verbal more effective than verbal only delivery, independent of retention interval, on both recall (d = 1.04), and transfer score (d = 1.76) • No main effect of retrieval practice on test scores. Lower rates of forgetting, but non-significant. • No interaction between practice conditions and lecture format
Karpicke *et al*. (2014)	Effect of retrieval practice on recall of science texts	3 experiments Age 9-11 years *N* = 94, 103, 89 1 elementary school, 4 classes US	***Neutral*** • Expt.1: no difference in scores between groups (authors suggest due to lack of support/guidance) ***Positive (but very small)*** • Expt.2: effect sizes small, but hint at a general advantage of concept map activity with less support (I.e., partially completed) relative to the condition that provided the most support (d = 0.12–0.17) ***Positive*** • Expt. 3: advantage of guided retrieval (retrieval using partially completed concept maps) over restudy, (d = 0.42)
McDaniel *et al*. (2011)	Effect of quiz frequency and placement on science test scores	2 experiments 8th grade 1 middle school US Expts 1/2a: *N* = 139 Expt2b: *N* = 148	***Positive, but relative to no retrieval rather than restudy*** • Quizzing produced significant learning benefits, with between 13% and 25% gains in performance on summative unit examination (Expt. 1) • Benefits of quizzing (relative to not quizzing) persisted on cumulative semester and end-of-year exams as well as end-of-unit exams (Expt.2) • Quiz placement: Review quizzing produced the greatest increases in exam performance (Expt.2)

McDermott *et al*. (2014)	Effect of quiz type on history and science test scores	2 experiments, same sample *N* = 141 *M* age = 12.85 1 middle school US	***Positive, but relative to no task rather than restudy*** • On the unit exams and on an end-of-semester exam. students performed better for information that had been quizzed than that not quizzed • The format of the quiz (multiple-choice or short-answer) did not need to match the format of the criterial test (e.g., unit exam) for this benefit to emerge
Ritchie *et al*. (2013)	Effect of retrieval practice (with or without mind-mapping) on geographical fact learning	2 experiments *N* = 109/209 Aged 8-12 years 1 primary school, 4/8 classes UK	***Positive*** • Overall: retrieval practice is more effective than concept mapping, and is not enhanced when concept mapping is added to it • Expt. 1: children in the retrieval practice group recalled significantly more facts than those in the non-retrieval practice group (d = 0.32, 95 % CI = -0.06, 0.70), but no effect of concept mapping • Expt. 2: main effect of retrieval practice (d = 0.43, 95 % CI = 0.14, 0.72)), no effect of concept mapping, and with results consistent at both 1 and 5 weeks later
Urhahne *et al*. (2013)	Effect of retrieval practice task type on science knowledge	*N* = 196 *M* age = 14.66 4 high schools Germany	***Positive, but relative to no task rather than restudy*** • Gap-fill and matching retrieval tasks were most effective in promoting knowledge acquisition, followed by multiple-choice tasks, and then no tasks at all (*η2 = .898)*
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Barenberg and Dutke (2019)	Effect of retrieval practice on comprehension accuracy and confidence in judgements	*N* = 98 Age 10-13 years 2 schools, 4 classes Germany	***No restudy condition.*** ***Positive compared to non-quizzing.*** • Students provided more correct answers in the testing condition than in the control condition (*d* = 0.29) • Students more confident in their answers in the testing condition (*d* = .70) • Confidence judgments were more accurate (effect on metacognitive monitoring)
Carpenter *et al*. (2009)	Effect of review/testing on recall of US history facts	*N* = 62 8th grade 1 school, 5 classes US	***Positive*** • Main effect of review method (within-subject): Highest scores for questions covered in retrieval condition (d = .32-.51) than Study or No Review items (latter two did not differ from each other) • Main effect of review group, and review method x review group interaction not significant
Dirkx *et al*. (2014)	Effect of testing on learning from principles and procedures from texts	*N* = 38 Age 15-16 years 1 school, 2 classes The Netherlands	***Positive*** • STST group outperformed SSSS group on the factual knowledge questions (*M* = 49.50%, SD = 22.59%; *M* = 21.67%, SD = 15.44%) (*d* = 1.44) • STST group outperformed SSSS group on the application questions (*M* = 60%, SD = 23.40%; *M* = 37.78%, SD = 22.64%) (*d* = 0.96)
Goossens *et al*. (2014a)	Effect of retrieval practice and learning context on vocabulary learning	*N* = 60 Aged 8-11 years 3 primary schools, The Netherlands	***Positive for retention*** Retention: • Children in the word pairs condition recalled more synonyms than children in the story condition (d = 0.40, 95 % CI = 0.11. 0.91) • Children recalled more words when learnt via retrieval practice compared to restudy (d = 0.52, 95 % CI = -1.04, -0.07) • No sig interaction
			Neutral for recognition Recognition • Children in the word pairs condition recognised more synonyms in the word pairs condition than the story condition (d = 0.73, 95 % CI = 0.21,1.25) • No effect of retrieval practice on recognition and no sig interaction
Jaeger *et al*. (2015)	Effect of retrieval practice on recall of information from texts	*N* = 69 8-10 years old 4 elementary schools Brazil	***Positive*** • Evidence for testing effect: children from the test group recognised an average of 17.4 (SD = 1.81) target facts (87%), whereas children from the restudy group recognised an average of 10.6 (SD = 2.98) target facts (53%) (d = 2.87, 95 % CI = 2.19, 3.55)

Karpicke *et al*. (2016)	Effect of retrieval practice on word recall	*N* = 88 Aged 9-12 years 1 school, 4 classes US	***Positive*** • Advantage of retrieval practice over repeated study in Expt. 1 (*d* = 0.48) and in Expt. 2 (*d* = 0.56) • Recognition was higher for retrieved vs. restudied words (*d* = .64) (Expt.3) • Effects were independent of individual differences (reading comprehension; processing speed)
Lipowski *et al*. (2014)	Effect of testing on word recall	*N* = 81 1st grade (6-8 years); 3rd grade (8-9 years) US	***Positive*** • Main effect of encoding condition, with greater performance on the final free-recall test in the test-plus-restudy condition compared with the restudy condition (d = 0.80, 95 % CI = 0.36, 1.24). • Same pattern for both 1st graders and 3rd graders
Nungester and Duchastel (1982)	Effect of testing on retention of history knowledge	*N* = 97 Age unclear ('seniors') (≈age 17-18) 1 secondary school, US	***Positive*** • Test condition resulted in better retention than either the restudy or the control conditions (restudy and control conditions not statistically different from each other) (d = 0.64, 95 % CI = -0.12, 0.86)

* High priority study identified for in-depth analysis; ^ = study included for more than one strategy.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in Table B3.4.

Table B3.4: Retrieval practice—quality of evidence assessment (based on the GRADE approach)

Strategy	Retrieval Practice (Compared to restudy or re-presentation)
Number of Studies	There are 21 studies in this area of which 16 provided evidence of retrieval practice compared to restudy or re-presentation specifically. Of this latter group, two were rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	All studies are randomised experiments.
Risk of bias	Our risk of bias assessments on the high-quality papers identified concerns with the randomisation, fidelity, and measurement for one study and with the selection of reported results for the other. Both had an overall rating of 'some concerns'. We judge, therefore, there to be at least one strong study in this area.
Inconsistency	**Result consistency**. The results were largely positive but with appreciable inconsistency. Of the 22 results from 16 studies[9] comparing retrieval practice to restudy/representation, 14 provided positive results, albeit with a large variation in effect sizes from very small (d = 0.12) to very large (d = 2.87). There were five neutral results and three negative results.
Indirectness	**Practice heterogeneity**. As we discuss at the outset and further in the discussion and questions section, retrieval practice encompasses a wide variety of techniques for eliciting responses. Moreover, the evidence in this area covered many subject areas. This variety adds confidence to our ability to make more general claims from this evidence. On the other hand, it reduces our ability to make claims about specific approaches to eliciting knowledge within retrieval practice. **Population, measure, and outcome heterogeneity**. There was good variety in the age range of students and subject areas examined, although there were fewer studies representing very young children. The vast majority of outcome measures were research-designed tests aligned to the targeted content, which was, in most cases, drawn from the regular curriculum. **Design and delivery**. This was an area with relatively low ecological validity. The key issue was that the vast majority of studies (18/21) were delivered by researchers using highly scripted, standardised procedures.

[9] We are counting Churches et al. (2020) as a single study in this count and have made a judgement as to whether studies that present multiple experiments and sub-analyses represent single or multiple results.

Imprecision	**Group sizes**. Most studies were small (n < 100) to medium (101 < n < 250) in scale. There was one large scale study in this area (n > 1000). Effect size estimates from high priority and medium-sized, medium priority studies reporting confidence intervals were: - Churches et al. (2020): the pooled effect estimate from retrieval practice RCTs was estimated as d = 0.28 (r = 0.14, 95% CI 0.06–0.23; d = 0.28; p = 0.001); - *Roediger et al. (2011): t(62) 6.61, d = 0.83); - Hanham et al. (2017): d = 1.37, 95% CI: 0.04, 1.65; and - Ritchie et al. (2013): Expt. 1, d = 0.32 (95% CI: -0.06, 0.70); Expt. 2, d = 0.43 (95% CI: 0.14, 0.72).
Publication bias	The majority of the small studies have positive effects compared to a more mixed picture for medium-sized studies. This suggests publication bias for smaller studies.
Other considerations	In this area we have compared retrieval practice specifically to restudy or re-presentation of material. This was a decision made prior to analysis. If we instead included all retrieval practice studies, this would include five further studies: one high priority, three medium-sized, and one smaller study—all of which provide positive results (compared to no retrieval practice condition).
Overall confidence	**Low (++)** Our confidence in the effect estimate is limited: the true effect may be substantially different from our estimate.
Confidence reasons	The reasons for the low confidence rating for this result were: - there was a large range of effects in this strategy group, with no *a priori* theoretical reason why this might be the case; - there were few high priority studies, and no larger medium-eligibility studies; - ecological validity was low: the vast majority of studies (18/21) were delivered by researchers using highly scripted, standardised procedures; - the vast majority of outcome measures were research-designed tests aligned to the targeted content; and - there are suggestions of publication bias for smaller studies.

Summary of findings for this strategy

Main finding. Overall, evidence suggests that retrieval practice has a moderate effect compared to restudy although there was high variability in this result. The evidence for the effectiveness of retrieval practice against a non-retrieval practice was more secure.

Estimated impact. Effect size estimates ranged from very small (d = 0.12) to very large (d =2.87). There were several negative results. The available evidence hints that moderate to large effect sizes are possible, but many instances have neutral or even negative results.

Confidence in impact estimate. Our confidence in this estimate is low, in particular due to the high inconsistency of results and low ecological validity. We have not formally assessed confidence in retrieval practice against a no-practice condition but believe that confidence in a moderate effect size for this would be low to moderate.

Heterogeneity/indirectness. There were two comparison conditions considered in this section. In many studies, the effect of retrieval practice (usually a form of quizzing) was compared against a non-quizzing condition. These studies were all positive and supported the conclusion that quizzing is an effective learning approach. As discussed at the outset of this section, however, the cognitive science principles for retrieval practice suggest that testing will outperform restudying conditions.

Other points. One of the negative results was from Hanham et al. (2017) in a retrieval practice test for learning content with high element interactivity (that is, connections and schematic organisation). The authors of this study theorised that it would be the case that the cognitive load (see next section) would adversely impact the effectiveness of the retrieval practice. This result, therefore, was in line with the overall theory. Our plan for analysis was to compare retrieval to all other forms of restudy (in their experiment, worked examples were used). We have listed this as a negative result rather than

changing the planned comparison *post hoc*. The value of retrieval for learning content with high element interactivity is an interesting question to which we return on the discussion and questions section.

Retrieval practice—overall evidence summary and conclusions

Summary of results

In this section, we have reviewed 21 studies focused on retrieval practice of which 16 tested retrieval practice against restudy or re-presentation of material. Our results for these are summarised in Table B3.5.

Table B3.5: Retrieval practice—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Retrieval practice (compared to restudy or re-presentation)	Twenty-one, of which three were graded as high priority.[1]	The overall evidence suggests that retrieval practice is an effective learning approach per se (i.e., against a no-treatment condition). Against restudy or representation of material, we judge there to be a positive effect overall, indicating moderate effect sizes.	A good range of learning areas was examined within the studies. The learning outcomes tended to be a factual recall or vocabulary learning although there were a small number of examples of learning with higher 'element interactivity', where elements needed to be connected as well as recalled.	Low (++)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Conclusions about strategies in this area

Retrieval practice and the testing effect

Our headline conclusions in this area are:

- Retrieval practice is highly relevant across the U.K. education system, for all learners and subjects.
- The findings in this area are mostly positive, suggesting moderate effect sizes, but there were an appreciable number of neutral or negative results.
- There was a good range of subjects and ages represented in the results. This suggests that the principle might have wide applicability across curriculum areas.
- One issue in this area has been the low ecological validity of the studies. The vast majority were designed and delivered by researchers, often in schools but outside of the classroom. Moreover, many interventions have been wholly scripted with a standardised procedure. This raises questions about whether 'real' teachers are likely to achieve the same results in more realistic conditions. The results from Churches et al. (2020), one of our high priority studies, suggest so but a firm conclusion is not possible based on the limited evidence we have in this area.
- We have focused specifically on studies testing retrieval practice against restudy or representation of material. However, we would note that re-study or re-presentation is likely to also result in learning. Using this as a comparator is perhaps a demanding test of the strategy. If we look at all 21 studies, this would add (with five studies with positive results) further weight of evidence to the conclusion that retrieval practice is effective.

Finally, we note that a recent systematic and meta-analytic review of testing, looking across all age ranges and a wide range of contexts, provides additional weight to these conclusions (Yang, Luo,

Vadillo, Yu, and Shanks, 2021). Yang et al. (2021) was published during the final stages of the write-up of the present review. This study estimated a medium overall effect of testing (quizzing: g = 0.50, CI: 0.44, 0.56) and also provides support for the use of different test formats and corrective feedback. As well as noting the main result here, we reference several of the findings from Yang et al. (2021) below where they provide further evidence in response to several questions we pose.

Evidence-informed discussion and questions

Principles and moderating factors

The testing effect

We began this section by linking retrieval practice to—broadly conceived—testing. Our definition of the area for the systematic review of the evidence centred on the finding (largely rooted in cognitive psychology) that testing can be more effective than restudying or re-presentation of the same material. Key ideas within this conception are, first, that a memory has a 'strength' and, over time, a memory's strength diminishes.[10] Many teachers in our interviews and questionnaires specifically referred to the 'Ebbinghaus Forgetting Curve' (see Murre and Dros, 2015) and the notion that students' ability to recall information drops dramatically within hours and the first 24 hours after learning and then more gradually, but still appreciably, over the following weeks and months. Second—and a response to this problem—is that learning requires some form of revision to counter the issue of forgetting. Literature and teacher survey responses frequently likened this idea to that of spaced practice, reviewed in a previous section. In theory, spacing the revision of material should counteract the forgetting curve and enable students to consolidate memory gradually into long-term memory. This is supported by neuroscience evidence about the processes by which retrieval practice facilitates learning by strengthening memory (Wiklund-Hörnqvist, Stillesjö, Andersson, Jonsson and Nyberg, 2021). The third connected idea here—the central idea for retrieval practice and the focus of our evidence review—is the claim that retrieval of information from memory will be more effective than restudy (such as re-reading material) or representation (for example, a teacher recap). This idea, sometimes called the 'testing effect', views testing as a 'desirable difficulty'; as we outline in the introduction to the evidence review in this area, while restudy can feel easier and students often perceive their learning as greater, the central idea of retrieval practice is that it is the strengthening of long-term memory (through the desirable difficulty of retrieving information from it) that is more important than the ease and vividness of bringing learning content into working memory through re-presentation or restudy. The evidence that we have reviewed supports this general idea: that retrieval practice is an effective learning approach *per se* (that is, against a no-treatment condition) *as well as* compared to restudy or representation of material, with the evidence suggesting a positive (and moderate) effect overall. With the limitations in the evidence-base and number of studies, our confidence in the estimated effect size was low but the evidence is fairly supportive of it being positive. Below, we explore several related ideas and facets of overall thinking around the testing effect, primarily drawing on authoritative reviews of the cognitive psychology literature.

What are the benefits of testing? What are the mechanisms for a testing effect? To what extent do these apply in realistic school classroom environments?

Above, we outline the central ideas of the testing effect. The practice review and basic cognitive science literature we have reviewed suggest that claims about possible benefits of testing extend far beyond the core ideas, such that (a) it is valuable to strengthen memory and (b) tends to be more effective than restudy. Claims made for the benefits of testing are wide ranging and impressive. The most vivid account of the benefits of testing are in Roediger et al. (2011) who describe 'ten benefits of testing and their applications to educational practice':

[10] For a concise introduction, see: Yan, V. (n.d.) 'Retrieval Strength vs. Storage Strength', Learning Scientists blog, guest post: https://www.learningscientists.org/blog/2016/5/10-1

1. The Testing Effect: Retrieval Aids Later Retention
2. Testing Identifies Gaps in Knowledge
3. Testing Causes Students to Learn More from the Next Study Episode
4. Testing Produces Better Organization of Knowledge
5. Testing Improves Transfer of Knowledge to New Contexts
6. Testing can Facilitate Retrieval of Material That was not Tested
7. Testing Improves Metacognitive Monitoring
8. Testing Prevents Interference from Prior Material when Learning New Material
9. Testing Provides Feedback to Instructors
10. Frequent Testing Encourages Students to Study

(Roediger et al., 2011, p.4)

This list of benefits is based on a review of the research literature, primarily from cognitive psychology. In our review, we have not been able to locate a sufficient number of applied scientific studies that provide robust tests of these in realistic classroom settings for school-age pupils. Therefore, this list is presented as a set of claims made about the benefits of testing that are yet to have been demonstrated in applied classroom trials with school-age pupils. Nonetheless, they provide a strong starting point for examining the mechanisms behind the testing effect.

The other area where investigation of principles and moderating factors is likely to be fruitful relates to storage strength and the distinction between retrieval and storage strength. This area of cognitive psychology distinguishes between learning and performance. In an informative blog summary, Yan (n.d.) uses the example of information that has high retrieval strength (for which one's test performance would be high if tested) but is likely to be forgotten in time, such as your current hotel room number, and compares this to a childhood phone number (low retrieval strength, high storage strength), a current phone number (high retrieval and storage strength), and a hotel room number from last year (low storage and retrieval strength). This connects to cognitive scientific literature examining the differences between learning and performance (Bjork and Bjork, 1992; Soderstrom and Bjork, 2015). There are a set of research questions here about how to distinguish these memory strengths and the principles and moderating factors for storage and retrieval strength. There are also practical questions apparent, such as how, through forms of assessment, teachers might distinguish between these. We return to some of the questions around implementation below.

One final idea relating to principles and moderating factors stems from teacher discussions of retrieval practice in our interview and questionnaire data. In addition to the core ideas (above), we note that many teachers framed the core principles of retrieval less in connection to the 'retrieve versus restudy' distinction but more in relation to the idea of the forgetting curve and the implications for curriculum sequencing with a general expectation that material should be revisited. Teachers were far more likely to talk about retrieval practice in terms of the need to space (Dunlosky and Rawson, 2015, p.75) and revisit learning than in terms of restudy or recap opportunities being replaced with tests. By way of example, one teacher responding to our questionnaire described the value of retrieval practice as follows:

> *'Retrieval practice, regular, is vital for preventing forgetting and keeping vocabulary etc. in LTM [long-term memory]. And spaced practice is crucial so it stays there. Before learning about these concepts, I often expected students to retain what they learned; now I understand I must give opportunities for students to remember what they've learned.'*

(Questionnaire response)

This idea about the need to constantly revisit the material seemed far more prominent in the teacher accounts in terms of a principle rather than the value of testing per se. Although, on the other hand, teachers' accounts of the strategies used for retrieval practice were strongly centred on low-stakes tests and quizzes.

Retrieval practice success, assessment, and feedback

What is the role of feedback in retrieval practice learning? Should error correction/feedback be built into retrieval practice activities? What are the implications of retrieval failure or errors in responses?

When one provides a test to a group of students, they will—to varying degrees—be able to (a) successfully retrieve some answers, (b) be unable to retrieve others, and (c) erroneously retrieve some answers (that is, surface misconceptions). This raises several questions:

- How does one determine optimal task difficulty?
- Is retrieval of misconceptions harmful?
- What is the value of providing feedback to give a second learning opportunity to fill gaps and tackle misconceptions?

The above considerations are related. If student performances reveal high rates of error, or low success, a teacher might decide to reduce the difficulty of the retrieval activity to keep the focus on content that has already been successfully learnt and focus on its consolidation in memory. Alternatively, a teacher might decide to include feedback and perhaps even opportunistically re-teach or clarify information following retrieval practice with a view to both consolidating existing learning and deepening and extending it. Several studies in our wider evidence-base (all medium priority) examined the additional value of feedback provision alongside retrieval practice. A short summary of these is provided below:

- **Damhuis et al. (2015)** was included in the main review, with a negative testing effect. They investigated the effect of added feedback on breadth and depth of vocabulary learning in Dutch kindergarten children (four to five years of age). They devised an intervention with a storybook reading of approximately 20 minutes followed by three experimental conditions: (a) repeated storybook reading, where the same story was read in the second and third week after the initial one, (b) repeated testing, where the children's target vocabulary was tested in the same weeks using a text-dependent picture task, and (c) repeated testing with feedback, similar to (b) but with feedback when words were responded to incorrectly. Based on this experiment, the authors *did not* find evidence for a testing effect. Their main conclusions were that repeated storybook reading had similar effectiveness for stimulating the breadth of vocabulary knowledge as testing and was better for stimulating deeper vocabulary knowledge. Of particular interest for the present question is that they also found that, for both breadth and depth of vocabulary, the effects of testing were conditional upon the provision of feedback.
- **Lipko-Speed et al. (2014)** also looked at the effect of testing with feedback but within science. In their study, a group of fifth-graders studied 20 science concepts and were subsequently divided into three groups: study only, test only, or test plus feedback. Following this, they were tested on the concepts. The findings showed that in the final test of difficult science concepts, students

performed best when they had been in the test-plus-feedback condition, but the authors argue that these two variables were largely independent (that is, did not interact with one another). In addition, the students, in general, had relatively low performance on this test and the authors argue it might be because they did not have enough time to study the concepts before being tested on them.

- **Kliegl, Abel, and Bäuml (2018)** similarly studied whether the testing effect was dependent on the type of test used as part of a retrieval activity but their comparison was between cued recall versus free recall and in the context of pre-school children. They also investigated the effect of immediate feedback. In their study consisting of four different experiments, groups of preschool children were asked to study pictures of fruits or animals and then divided into groups that re-studied the material with different types of tests and with the use of immediate feedback. Their study found that a free recall final test resulted in no testing effect, regardless of which type of test the children had used as part of their retrieval. However, if a cued-recall test was used in the retrieval and final test, a testing effect was found, and this rose significantly when the children had had immediate feedback during the retrieval practice test.
- **Leggett et al. (2018**) explored whether retrieval practice produces benefits for all learners. In addition, they explored whether teachers could enhance retrieval practice by adjusting its difficulty using hints. Their experiment involved a group of Year 9 geography students listening to some material and then reviewing some of the material by reading, some by answering questions and then performing a feedback task, and some not at all. Furthermore, a hint sentence could be given within each reviewed item. They were tested after one week on all the material. Study findings showed that 'retrieval practice with feedback' improved retention of facts more than did reading. This was true regardless of students' practice test performance, enjoyment of the activity, or belief in its effectiveness' (p.765). This, the authors argue, suggests that retrieval practice is 'relatively effective for all learners, at least when it is preceded by some prior exposure to the material and followed by appropriate feedback' (p.765). Based on this, they did not find a consistent effect of hints.

These studies suggest benefits of feedback may be connected to retrieval practice. Lipko-Speed et al. (2014) suggest that these might be considered to be separate learning approaches working independently. Yet, Damhuis et al. (2015) conclude that the benefits of testing were conditional on the provision of feedback. Of interest for the present question is the result of Leggett et al. (2018) that retrieval practice was effective 'regardless of students practice test performance'. There were other studies, like Leggett et al. (2018), that examined whether hints and cues as part of feedback could support and enhance retrieval practice, summarised as follows:

- **van den Broek et al. (2019)** explored hints as part of feedback, investigating 'whether retrieval practice with hints feedback is more efficient for recall several days after practice than retrieval practice with show-answer feedback' as hints essentially provide another opportunity for retrieval. They did not find any significant difference between retrieval practice with orthographic hints feedback and retrieval practice with show-answer feedback and therefore conclude, 'We found no clear benefits of hints feedback that created an extra retrieval opportunity compared to show-answer feedback (p.11).'
- **Mateo et al. (2020)** explored the effect of cues on memory retrieval, comparing visual and verbal reminders. Following a farm school trip, a sample of preschool children (50 to 74 months of age) were measured for their initial memory of the trip and again after having a reminder cuing procedure in which the event was reviewed across eight sessions using either visual or verbal reminders (and a control group not receiving reminders). They found that the experimental groups scored higher on the post-procedure test than the control group, demonstrating the effect of cues

on memory. However, they did not find any difference between the two types of cues and no improvement compared to the control group after six months.

Results such as these and the general lack of applied evidence make firm answers difficult for questions about how much 'difficulty is desirable' or whether feedback might provide a solution for unsuccessful retrieval (as well as another potential retrieval practice opportunity). On the other hand, corrective feedback was identified as an effective enhancement to retrieval practice in the comprehensive review and meta-analysis of Yang et al. (2021); they estimate a positive overall effect for testing with corrective feedback compared to testing without feedback (g = 0.54 compared to 0.37). While this result was across all age ranges and contexts, and the majority of studies involving university students, this provides strong support for the addition of effective feedback against the more mixed results reported above.

Further complexity arises when looking across cognitive science concepts. For example, many teachers and practice review sources firmly linked retrieval practice and spaced practice. Both are forms of desirable difficulty for which the effects may not be entirely additive, as explained in a point by Putnam and Roediger:

> *'Combining retrieval practice with spacing creates an interesting situation, however, because retrieval practice is more effective when it is successful, but increasing the gaps between retrieval practice opportunities makes retrieval more difficult (less successful). Thus, the positive effects of spacing and retrieval practice may be working against one another, at least when no feedback is provided on the test.'*
>
> *(Putnam and Roediger, 2018, p. 177)*

One final question we have in this area relates to the fact that many teachers who we spoke to reported that they used retrieval practice as a part of their formative assessment practice to identify what students do not know (and misconceptions) as much as strengthening what they do. One of the teachers we interviewed (Interviewee 13) connected a wide range of ideas about student misconceptions, assessment, and data in their discussion of retrieval practice, as follows:

> ***On misconceptions**: 'The issues that you have around multiple choice are around making sure that teachers know how to write a multiple-choice question. And making sure that teachers understand that you've got to put misconceptions into the choices of answers in order to surface them … What are you going to take out of the lesson?' … And I think one of the things that's interesting in our school is that previously, we saw any kind of retrieval practices as, "Can they do the thing I've taught?", but what we've trained the teachers in now is not to look at it like that. We've trained them in the formative assessment side of things in terms of the function of it is to spot what the children can't do, and therefore what to do. So teachers in our school now talk about finding out what can't be done and intentionally surfacing those misconceptions. Because that's actually more important to us than what they can do and what they do know.'*
>
> ***On data and tracking**: 'Do you collect that data? … you know, the whole point of it isn't that we have millions of trackers. And that's become a thing in the Ofsted framework, which is quite interesting, you know; they're no longer interested in looking at internal tracking systems in schools. So given that the research says it's*

not about creating loads of trackers and Ofsted aren't really that interested in trackers and we don't want to increase teachers' workload … what do we collect data on? What do we not collect data on? If we're collecting data on it, then we have to do something with it … a leader has to then have some way of getting some sense at key points of the key stuff that the kids can't do. Because they then have to help teachers plan patterns that have emerged across the groups.'

(Selected quotations from Interviewee 13)

We stress that we do not present this particular account to justify one way of understanding retrieval practice over another but as a way of posing questions and probing issues around the principles of the approach. What is evident in Interviewee 13's perspective is that retrieval practice, as a form of testing, can potentially be framed as much as a formative assessment approach (that is, supporting teachers to identify and address gaps in teaching and learning) as one centrally focused on revision (the retrieval of information from long term memory to consolidate learning). We do not know from our data what is emphasised within actual teacher practice and whether revision and new learning should be thought about and practiced in combination. Moreover, only a few applied studies examined misconceptions in this area. One exception was **Chang's (2010)** study of 208 11th- and 12th-grade students in Taiwan. The study examined the use of correct and incorrect concepts in multiple-choice and true/false statements on a test. Their results suggest the danger of misconceptions arising from incorrect test items and recommend providing students with the correct answers immediately after the test. While the applied evidence is too limited to support the point, it would seem plausible that retrieval tests that might surface (and potentially reinforce) misconceptions will benefit from the provision of feedback and re-teaching more than tests designed for high retrieval success. One final, related consideration raised in one of our interviews is considering the emotional (and self-efficacy) aspects of the success or not of retrieval. They reflected that they had less success with retrieval practice for this reason:

'I feel a bit overloaded, because there is so much. Happy I have done some reading, but now I need to focus on this and give it some time. Because it is not all good news. One child who gets the answer wrong, that is a quick failure. There is quick feedback in quizzing. But it can reiterate failure in some children and I hadn't expected that at all, so it is about being aware of how it works on the ground.'

(Interviewee 5)

Variation in the practice or teaching and learning context

Format of retrieval tests

Does the format of the test—multiple-choice, short-answer, long answer, cloze, rapid generation of responses (such as writing lists)—affect retrieval practice effects and efficiency?

When discussing the use of tests for retrieval practice, one consideration is that testing has, for many, negative connotations and is linked to high stakes standardised examinations. Popular scientific accounts of retrieval practice encountered in the practice review often stressed that when using tests for retrieval practice, it is important that these are low-stakes or no-stakes (for example, Jones, 2019). One study in our wider evidence supported this suggestion, **Mok and Lan Chan (2016)** found that for students with high test anxiety, summary writing was a more effective retrieval practice approach

than testing. From the perspective of the basic scientific theory, retrieval practice is simply the idea that memory accessed in long-term memory is strengthened. The specific teaching and learning strategies used to achieve this has not (as far as we can tell) been a major focus of the research. However, there were a small number of examples in our wider evidence (for example, medium priority studies not included in the main strategy systematic review group) looking at the effect of test format on learning from retrieval practice. We briefly summarise these studies below:

- **McDermott et al. (2014)** investigated the benefits of multiple-choice quizzes versus short-answer quizzes on later exam performance. Their study consisted of five experiments, conducted in a seventh-grade science classroom and a high school history classroom. They found that quizzing aided performance, regardless of the type of quiz and whether it matched the format of the final test. However, their findings did not provide a clear answer as to which of the quizzing formats they examined was best. Nevertheless, they argue that 'What is clear is that both types of quizzes give benefits, and a conservative conclusion across our experiments is that the benefits are roughly equivalent from the two types of test' (p.15).
- **Duchastel's (1981)** study looked at the effect of different types of tests on later retention. In his study, 57 secondary school students studied a brief history text before taking a short-answer test, a multiple-choice test, or a full free-recall test. Two weeks later, all participants were given a retention test. Duchastel found that the testing effect was evident in the case of the initial short-answer test, but not in the case of either of the other two tests. He thus argues that the testing effect can be strongly influenced by the type of test adopted. However, he also acknowledges that the retention test was closer in format to the initial short-answer test and that some of its impact could have been due to a practice effect.
- **Duchastel and Nungester (1982)** draws on this earlier work of Duchastel (1981) but differs from it; their retention test was of a similar format to both testing types (short recall and multiple choice). Following the learning of a brief text, participating students received either a short-answer test, a multiple-choice test, or no test at all. Two weeks later, all students received an unannounced retention test, which had both short-answer and multiple-choice questions. Findings showed a strong testing effect from both experimental conditions and that the initial short-answer test did not lead to better retention than the initial multiple-choice test. They argue that the results from the previous study (Duchastel, 1981) may in part have derived from the 'practice effect' and that the testing effect may therefore involve both a consolidation and a practice component.

Overall, the evidence suggests that different test formats are likely to influence the effect but that content-relevant and practicable test formats in a variety of forms can be used. This is further supported by Yang et al. (2021) who found variation in effects by test format; *g* effect size ranging from 0.91 to 0.24 for test formats of matching (0.91), fill-in-the-blank (0.64), short answer (0.57), multiple choice (0.31), cued recall (0.24) and free recall (0.34). They also found that consistent test formats across retrieval practice attempts were associated with larger effect (g = 0.53) than inconsistent formats (g = 0.40).

Does learning from retrieval practice transfer to other tests? How limited is the learning to the specific content and retrieval approach? Does varying the content or retrieval approach promote transfer?

One issue touched on by Duchastel (1981) and Duchastel and Nungester (1982) is the issue of the 'practice effect' and the related question about transfer. We identified two studies that considered

issues relating to the practice effect and the extent to which tested material will 'transfer' to other tests, item types, or related learning areas:

- On the question of consolidation versus practice, **McDaniel et al. (2013)** investigated 'whether learning from quizzing arises from memorization of answers or fosters more complete understanding of the quizzed content'. They argue that previous studies have often used the same material in the retrieval tests and the final tests and that the benefits of retrieval in these cases could simply be due to practicing the answers. Their study, however, found that 'spaced testing (quizzing) procedure, along with feedback, can promote learning that is deeper than just retaining a particular answer' and thus does 'enhance knowledge, which can be flexibly used for different test items appearing on later exams' (p. 368).
- **Rohrer et al. (2010)** also looked at the question of transfer. They conducted two experiments with fourth- or fifth-grade students who 'learned to assign regions or cities to map locations and returned 1 day later for 2 kinds of final tests'. One of the tests required the same task seen during the learning session whereas the other consisted of novel, more challenging questions. The authors found a testing effect for both kinds of final tests and, furthermore, that the testing effect was no smaller (actually slightly higher) for the final test requiring transfer.

In their interview and questionnaire responses (see the Interleaving and Space Practice section for examples), several teachers indicated that they were deliberately trying to vary content and retrieval practice approaches to promote transfer and elaboration of learning (also see Pan and Agarwal, 2018).

Retrieval practice, element interactivity, and cognitive load

Does retrieval practice work in all subjects and topics? Does retrieval practice work for learning with high complexity, subtlety, or element interactivity as well as factual recall?

The range of subjects covered in our main results suggests that retrieval practice is used in some form across most subjects (also see Yang et al., 2021*).* A more challenging question touched on in our main results is whether the applicability of retrieval practice is limited to learning content with low element interactivity (that is, disconnected factual recall knowledge, suitable for rote learning). One study in our main evidence, **Hanham, Leahy and Sweller (2017)**,[11] examined high versus low element interactivity knowledge retrieval. In our summary of this work in the main results, we reported a positive testing effect for lower element interactivity materials and a reverse testing effect (for immediate, but less so for delayed, tests) for high element interactivity materials. We now look more closely at this study.

- **Hanham, Leahy and Sweller (2017)** discussed the interaction between retrieval practice and the complexity of information. They examined whether the testing effect was evident under low or high element interactivity/complexity conditions based on the cognitive load theory. Element interactivity can both be related to the nature of the information and the expertise of learners: if the learners' increase in expertise, the element interactivity of a task decreases (as learners can draw on organised information in long-term memory). Based on six experiments with differing levels of material complexity and different age groups, they found that 'tests can have an important role in facilitating learning, but that faciliatory effect only occurs when the material being learned is low in element interactivity. Conversely, if it is high in element interactivity, a test can inhibit rather than facilitate learning' (p.279).

[11] Rated medium eligibility and included in the main results.

One other relevant study was Agarwal (2019)[12] who looked at whether building factual knowledge via retrieval practice could enhance students' higher-order learning. Her results indicate that it did not. Nonetheless, it does suggest that retrieval of higher order items was beneficial. The performance in tests was greater when the retrieval practice format matched the final test format: 'fact quizzes enhanced final fact test performance, and higher-order quizzes enhanced final higher order test performance' (p.202). She also found that mixed quizzes—that included both fact and higher-order questions—increased higher-order test performance more than fact quizzes and slightly more than higher-order quizzes, providing support for mixed quizzes.

Yang et al., in their systematic review and meta-analysis of retrieval practice across a wide variety of ages and settings, concluded that testing can benefit conceptual learning (g = 0.64) and the application of knowledge for problem solving (g = 0.45) as well as factual recall (g = 0.52*).* As a high-level meta-analysis, with few details provided about how 'Concept Learning' was operationalised, this provides encouraging but not conclusive evidence that retrieval practice can be beneficial for more complex learning. Considering this result alongside the two studies reported immediately above, we are not able to draw any firm conclusions about whether retrieval practice is suited for higher-order or high element interactivity learning or the extent to which learning content at different levels might transfer to higher or lower levels.

As we note in our main results on retrieval practices, study learning outcomes tended to be factual recall or vocabulary learning with only a small number of examples of learning with higher element interactivity. This was also reflected in teacher perspectives in our practice review where their descriptions of retrieval practice tended to focus on vocabulary learning or retrieval in maths, such as times tables facts. One teacher commented in our questionnaire, 'I think it is a shame that retrieval is seen so dogmatically as quizzing and planned questions rather than dialogic and generative questioning and going back over things to make sense of new things.' This general issue was also reflected in some of the challenges teachers spoke about:

> *'I tried using flash-cards to teach the definitions of word classes (nouns, adverbs etc.) and the pupils could repeat back the definitions parrot-style but then, if I showed the pupils a word, they still found it difficult to identify its word class.'*
>
> *'Retrieval practice is hard when kids are doing concept-based work ... The students struggle to memorise definitions verbatim and get bogged down with whether they get the mark if they've missed articles or pronouns and don't understand why if they have the overall meaning need to learn it word for word.'*

Many others were more confident about the use of retrieval practice to develop higher-order knowledge. One interviewee, for example, described their overall approach as—

> *'a kind of overall strategy. I suppose it is just regular memory retrieval and spaced practice are the ones we use most and then we kind of step up learning with the students, so they're gradually working up to things which are more challenging, gradually deeper knowledge, so we start with the basic knowledge and then we gradually move up to the evaluation application skills with them in terms of deeper processing ... so that rather than just saying "right you have learned this, revise this*

[12] Rated high eligibility but judged to be insufficiently similar to other studies to be assessed in a group.

and move on", we are constantly moving back, saying "right, now we are revising what we did last year or before Christmas".'

(Interviewee 2)

Implementation

What different retrieval practice activities or approaches are there and what are the challenges of implementing them?

In this final section, we look at some of the different retrieval activities, the approaches used, and the challenges of implementing them. We note that there are many practitioner-facing accounts of retrieval practice strategies within sources in our practice review (see for example Jones, 2018; Agarwal, 2021).[13] Here, we mostly draw on this practice-facing literature and the teacher perspectives from the interviews and questionnaires. As we note in the main evidence review, the vast majority of retrieval practice studies were designed and delivered by researchers, often in schools but outside of the classroom. Moreover, many interventions have been wholly scripted with a standardised procedure. This raises questions about whether 'real' teachers are likely to achieve the same results in more realistic conditions.

Retrieval activities

Teachers' descriptions of retrieval practice from our interviews and the questionnaire revealed that many use low-stakes quizzes, especially as lesson 'starters' for retrieval practice. Many were doing this in most lessons on a daily basis. Thus, retrieval practice appears to be a large part of many teachers' overall pedagogical approach:

'We use a lot more testing, but low stakes testing. We use mini whiteboards, we use red cards, we use quizzing, we ask students to talk things through before sharing them as a class, we use homework to help students to have to recall information. At the start of lessons, we have them answer questions as soon as they arrive, just three or four. And those are structured on what they learned yesterday, the previous week, the previous month, that sort of thing.'

(Interviewee 13)

A large number of retrieval activities were discussed, including:

- the use of knowledge organisers to rehearse key learning points;
- retrieval grids linked to units with questions to revisit covered content;
- 'starter brain dumps in the form of mind maps on last lesson content and compare to content from a few weeks back; I then encourage them to do this process regularly at home [too]';
- 'quizzes in history using colour cards to choose multiple choice answers';
- 'labelling diagrams with gradual reduction in information provided over time';
- 'creation of own flashcards using dual coding principles';
- 'retrieval roulette starters';
- 'these could be diagrams to label, questions to answer, gap fills to complete that are completed and self-marked by the students to encourage hard thinking and resilience (the expectation is that students will at least have a go)';
- 'true or false, multiple choice, cloze procedure and finish the sentence; and

[13] Resources, including free practice guides available from http://www.retrievalpractice.org

- 'Bjork's confidence weighted triangles'.

A small number of studies in our wider evidence compared retrieval practice approaches. A short summary of the focus and results of these is as follows:

- **Frits et al. (2018)** studied the effect of drawing pictures, writing down propositions, and muttering during memorization to remember concepts and relations between concepts. Their study found evidence that drawing led to better reproduction of concepts than writing during memorisation and that muttering did not make any difference compared to silently memorising.
- **Urhahne et al. (2013)** compared learning with gap-fill and matching tasks, learning with multiple-choice tasks, and learning only from text and figures without any additional tasks. They found that adopting a 'gap-fill and matching tasks' approach was the most effective for acquiring knowledge. This was followed by the multiple-choice tasks and, finally, by the no tasks at all.

Retrieval organisation and timing

As noted above, many teachers reported using retrieval practice within lesson starters. Many described their system for what was retrieved, especially in relation to the spacing of the content included in the retrieval practice: one commented, 'Regular use of "do nows" is proving effective in encouraging students to recall recent and previous learning; the activity will include a question from a topic learnt "last lesson, last week, last month" and a "skill for today".' Many teachers mentioned this kind of organisation ('last lesson, last week, last month'). One described this as a 'goldfish, dog, elephant starter: last lesson, last unit, last year.'

One interviewee described their use of the Leitner system for flashcards retrieval:

> *'So you write down what you learn on flashcards and the next day you test yourself. The ones you could remember go into the Thursday. You do one the next day and the ones you can remember you save. Its on YouTube. You make little flashcards and you test yourself everyday and by testing yourself everyday, it stays in your memory for longer. So that's where we've got to at the moment.'*
>
> *(Interviewee 9)*

More generally:

> *'The way we work our curriculum in my department, they'll do an assessment, they do regular formative assessments as they go, but they do end-of-topic summative assessment and then we build back things, so we will always go back. So now we will do an accumulative assessment, if you like, so we kind of go back to the beginning of the year, so we kind of dip in back to things that they have learned previously and we space it out.'*
>
> *(Interviewee 2)*

Some teachers described how they organised retrieval prospectively within lessons. For example, one (Interviewee 4) described how they end the lesson by discussing the most important piece of information and they then use this as a basis for a quiz in the next lesson. Another described student involvement in writing down vocabulary lists and putting these in envelopes for use in the future.

There was some discussion of organisation of retrieval in terms of planning. There were numerous mentions of planning retrieval into schemes of learning, in some cases this being systematised as a whole-school approach expected by school leaders:

'As a school, retrieval practice is the most embedded, with departmental leaders required to include retrieval opportunities within their curriculum planning. This has removed the need for extended revision time in the run up to mock exams and formal assessments. The language is used by students and teachers and they are used to the concept.'

(Questionnaire response)

Ease of implementation

Retrieval practice was an area where very few teachers discussed implementation challenges. Instead, issues raised were along the same lines as those for spaced practice, about the pressures of time to cover the curriculum and whether there was sufficient time to include large amounts of practice rather than covering new material. The only issue raised was by a school leader on how teachers could be best supported to use retrieval practice in connection with formative assessment:

***On teaching and feedback during retrieval practice**: 'How do we train a teacher to react in the lesson? You know, is it that they react in the moment? And you know, if a teacher diagnosis from the multiple-choice question [reveals] that … the kids can't do question nine … the teacher then has to make a decision? Can I tackle that now? Do I have time? Should I do that tomorrow?'*

(Interviewee 13)

Comments about implementation were highly positive, with example comments as follows:

'Retrieval practice is easiest to implement and the one students latch onto first. We start each lesson with a "do now" retrieval activity. These questions are often interleaved (questions from last lesson, last week, last term etc.) so also have the advantage of including the spacing effect. Students use retrieval practice for their own revision—Quizlet etc. very popular.'

'Retrieval is easy to plan, widely shared, and resources last … Retrieval practice every lesson with my starter … I use spaced retrieval on homework tasks and throw in exit tickets from past topics too.'

'[We use] retrieval practice as it is the easiest to explain and implement. We use implemented retrieval (based on the forgetting curve) of key fluency facts at the start of each lesson. We hope to space some practice of reasoning and problem-solving questions once this is embedded.'

Many (but not all) discussed retrieval as something students enjoy:

'Retrieval practice is working well and the learners enjoy the quizzes. We have canvas and an online VLE so they enjoy being able to get the results quickly. Also they enjoy the mastery paths that they can access following the quizzes.'

'Starting lessons with a five-question starter to interleave the learning has been particularly effective, not just in terms of retrieval but also for student confidence and self-esteem. This increase in confidence drives a positive cycle of engagement and further learning.'

Does retrieval practice work for pupils of all ages and prior attainment levels? For which groups and in what circumstances might retrieval practice be harmful?

In our main results section, as well as studies finding a positive impact of retrieval practice there were also several that found a negative impact. The negative impacts were found with primary age children (age 4 to 12). However, the evidence is not sufficient to draw conclusions by subgroup: these studies raise questions about whether these negative results represent usual variation in outcomes, weaknesses of the intervention in the specific studies, or differences in the applicability of retrieval practice by student age or learning outcome. With regards to ability levels, recent evidence from Jonsson et al. (2020) provide both behavioural and brain imaging evidence that supports the view that retrieval practice is effective for all, when dividing their sample of upper-secondary students into low-, medium- and high-cognitive-ability groups.

The only evidence we have on suitability across pupil ages and characteristics in our database comes from teacher interview and questionnaire responses. When discussing the students that benefit, most of the teachers we spoke to felt that retrieval in some form was suitable for all students. Some said that it was particularly suitable for students with special educational needs, some felt it was particularly helpful for higher-attaining students, some for lower-attaining ones, but, overall, there was no one pupil group identified for whom retrieval practice is more or less applicable.

> *'I have found lower-attaining students or students with less prior knowledge find retrieval tasks very difficult, even when they relate to very recent knowledge. This is probably because they know and remember little, so they find retrieval tasks very challenging and demotivating.'*

> *'Retrieval practice/low stakes testing works with all pupils but has most impact on low-ability, lacking-in-confidence pupils.'*

Final thoughts on this strategy area

In our systematic review of classroom trials, we concluded that retrieval practice was an effective learning strategy *per se.* The evidence supported the view that it tends to be more effective than restudying or representation. There were suggestions of moderate effect sizes, but our confidence was low in a particular estimate of an effect size, which we judge to be highly likely to be positive. A good range of subjects and ages were represented in the results, suggesting that the principle might have wide applicability across curriculum areas. Teachers we spoke to were highly positive about retrieval practice. One of our interviewees summarised their view on retrieval practice as follows:

> *'There are some … strategies that I am completely sold on; I think that I would never not incorporate retrieval practice and give students an opportunity to remember.'*
>
> *(Interviewee 4)*

The overall evidence is therefore positive for retrieval practice. One area of concern in the main results was the low ecological validity of the studies in the review: many interventions were wholly scripted with a standardised procedure. We have also raised several challenges around questions related to the complexity of learning suitable for retrieval practice, misconceptions, and retrieval practice success and the links between retrieval practice, feedback, and classroom assessment.

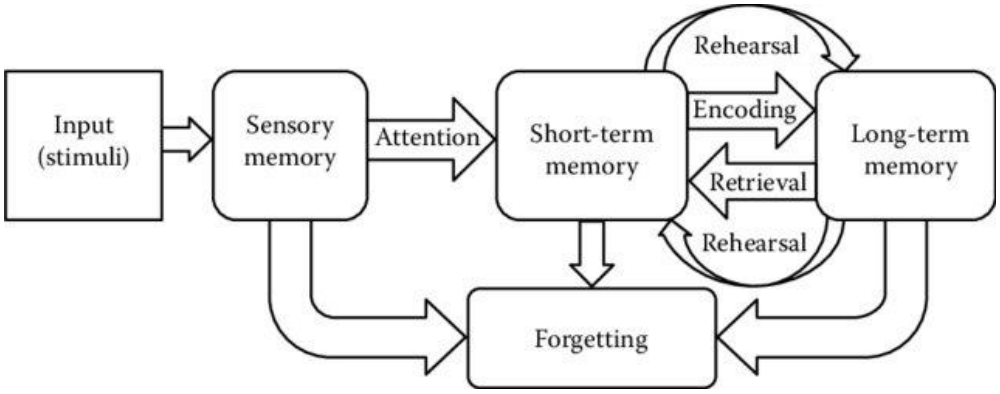
B4. Managing cognitive load

Overview of area

Definitions

Modern accounts of human memory, stemming from seminal work such as Baddeley and Hitch (1974) and Atkinson and Shiffrin (1968), distinguish sensory, working, and long-term memory. This separation of working memory from long-term memory is a widely accepted theory, held to be one of the most robust frameworks for understanding the mind in cognitive science (Didau and Rose, 2016). As described by Cowan (2014), working memory is 'the small amount of information that can be held in mind and used in the execution of cognitive tasks, in contrast with long-term memory, the vast amount of information saved in one's life' (p2). In a later section, we examine memory in more detail and, in particular, the role of visuo-spatial and phonological channels. For now, we focus on a central educational problem: the limited capacity of the working memory. This might be summarised as follows (for detail, see Cowan, 2014). First, the senses register input from the environment. The vast majority of this information is filtered out, with only a fraction of the available information attended to. This information then enters working memory (or short-term memory). From here, the person—through attention and rehearsal (but in some cases without either)—can encode the information into long-term memory. Long-term memory enables the individual to draw upon long-term memories in new situations, which can subsequently be revised, developed, and combined with new, incoming information. As we saw in an earlier section, retrieving information from the long-term memory store will strengthen and consolidate the long-term memory.

This account is depicted in the diagram below, from Amin and Malik (2014, p.221) based on Atkinson and Shiffrin (1968).



Source: Amin and Malik (2014, p.221) based on Atkinson and Shiffrin (1968).

The educational problem stemming from this overall 'architecture' of the mind is that while long-term memory is held to be practically limitless, working memory has a limited capacity. Specifically, working memory it is thought to be limited to hold around three to five 'chunks' of information at any one time (Didau and Rose, 2016, p.43), although what constitutes a chunk is not entirely straightforward and depends on the ability to group material into schema-like structures and draw on long term memory to prevent the need to hold information in working memory (see Smith, n.d., for a short discussion of this question). Overly detailed or complex presentation of information or problem spaces can easily overwhelm or 'overload' the working memory. Educational psychologists and cognitive scientists

make the distinction between (1) *intrinsic load* on the working memory, determined by the nature of the information to be learnt (for example, the difficulty of a task), (2) *extraneous load*, which is caused by unnecessary or distracting information that is not essential the learning of the target information, and (3) *germane load*, the load specifically used for developing knowledge (via the construction or alteration of schemas) in the long-term memory. Managing cognitive load is not a task of *minimising* cognitive load but rather *optimising* it. Theoretically, for optimal learning, educators should seek to minimise extraneous load while maximising intrinsic (or germane) load while not exceeding working memory capacity (see Lovell, 2020, for a concise, accessible, and educationally focused account). Evidence also shows that there is substantial variation in working memory capacity between individuals (Cowan, 2016; von Bastian and Oberaurer, 2014); effective cognitive load management will therefore require consideration of specific pupils and pupil groups and is likely to pose a greater challenge when teaching a full (and especially a mixed-ability) class.

Exceeding working memory capacity is a particular issue in relation to 'problem solving' in education, where learners are typically presented with a large amount of information, in quantity or complexity, and asked to successfully identify target information or follow (or sometimes discover) a series of problem-solving steps (Sweller, 1988). As a result, students, especially for those with limited prior knowledge, often struggle to 'navigate' through this problem space, working memory is overwhelmed, and learning is impaired.

In response to the finite capacity of students' working memory, there are educational strategies—examined in this section—which seek to reduce or optimise the load on working memory. Here we are going to review the evidence in three strategy areas:

1. using worked examples to support learners to solve complex problems and to apply and develop knowledge; in some cases, incorrect or incomplete worked examples are used to provide partial support, often with a view to gradually reduce or 'fade' the level of support provided;
2. providing 'scaffolding' guidance or other forms of support such as prompts, cues, or targeted instructions to reduce cognitive load and help learners navigate the demands of the task; and
3. conducting problem solving with a collaborative context in which learners can provide support and share the task's demands.

There is a great deal of variation in and around these strategies and other relevant considerations such as individual learner prior ability or knowledge and working memory capacity, and the effect of emotions such as anxiety on working memory. We also note there are connections elsewhere—in particular to our review of schema-based instruction strategies (see next section). Numerous studies described their approach as 'schema-based instruction'. Several of these are included in this section where there was a specific focus on reducing or optimising working memory demands. Another key link is to multimedia strategies, including strategies that 'dual code' information. Working memory is fundamental to learning. While making a clear-cut distinction is challenging, here we are focusing specifically on strategies that try to reduce or optimise the load on working memory.

Overview of the evidence-base

Table B4.1: Managing cognitive load—overview of study priority ratings

Priority level	Overall rating	Ecological validity	Relevance and definition for focus CS practices	Added value to evidence-base
High	7	8	31	31
Medium	84	71	106	88
Low	59	73	15	33

The review study database contained a large number of studies in the Managing Cognitive Load category. Of these, 93 were graded as being of sufficient ecological validity, relevance, and value for inclusion within this analysis of the evidence (high and medium). There was, however, large heterogeneity in study foci so after grouping studies testing similar approaches, we identified 45 studies for analysis in the main results, which we organised into three groups (a small number of studies were included in more than one category):

- **worked examples** (22 studies, of which four are graded as high); a subset of seven papers (of which two were graded as high priority and thereby identified for in-depth analysis) concerned the use of incorrect or incomplete worked examples;
- **scaffolds, guidance, and schema-based instruction** (16 studies, two graded as high priority); and
- **collaborative problem solving** (nine studies, of which one was graded as high priority).

Seven of the 45 studies selected for analysis scored highly across our criteria and were identified as *potentially* providing strong evidence in this area. We report selected evidence from the remaining 48 studies in the discussion section. While we did not judge the weight of evidence in each area to be sufficient for a dedicated analysis, numerous significant results add to the evidence and theory and signal the way for further research.

Wider evidence in this area looks at a range of issues including working memory improvement intervention, determinants of cognitive load including element interactivity, split attention, productive failure, anxiety as a factor impairing working memory, faded support, timing and sequencing, chunking and integration of material, and student differences.

Main findings

Strategy 5: Worked examples

Concise definition

Worked examples involves providing students with step-by-step, or part-by-part, demonstration of a task that makes clear the required product (answers or output) and the process of completing the task.

Full definition and description

Worked examples involves providing students with step-by-step, or part-by-part, demonstration of a task that makes clear the required product (answers or output) *and* the process of completing the task. Worked examples are often accompanied by explanatory notes, definitions, or reminders,

sometimes with prompts for reflection or self-explanation. Teachers also often 'model' the process through demonstrating the creation of a worked example with learners (partially or in full). In some cases, incorrect or incomplete worked examples are used to provide partial support, often with a view to gradually reducing or 'fading' the level of support provided.

Selected examples

Examples of this strategy from our database.

- Worked examples of geometry problems are discussed in Youssef-Shalala et al. (2014) who provided problems in pairs with a 'study one, solve one' approach; the first problem in each pair was a worked example, with solution and instructions, and the second left for the student to solve. Problem pairs were used in other studies, such as Retnowati, Ayres and Sweller (2016) and Retnowati, Ayres and Sweller (2010). Geometry worked examples included the diagram (for example, with a missing angle), a sum calculating the answer, and a stated theorem (such as '$x^o = 180^o - 54^o - 126^o$'—adjacent angles on a straight line sum to 180^o).
- Many worked examples (for example, Booth 2015) included instructions with 'write' icons (for example, 'your turn' for problem pair to solve or 'rewrite the question correctly' in the worked example). Many included arrows to prompt, explain, or emphasise key aspects of the problem.
- Several studies (for example, Kyun, 2009) made use of computer-based problems where students had to work though a series of steps.

Evidence for this approach

There were 22 studies relating to the use of worked examples. Of these, four were graded as high relevance and quality. We also identified a subset of seven papers (two rated high) within these concerned with incorrect or incomplete worked examples. We have separated these into a separate evidence table but will consider the studies as an overall group. Full details of all medium and high studies in the analysis are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed for the worked examples strategy are characterised as follows:

- **Pupil age and characteristics**. The vast majority of studies in the area (20/22) were for secondary age students (age 11 to 18) with a good spread of studies between age 11 and 16. There were very few studies of primary-age students, the exceptions being Van Loon-Hillen et al. (2012: fourth grade, age 9–10, n = 45) and Yang et al. (2016: third grade, age 8–9, n = 109). Primary age children are therefore poorly represented in our data and early years children not at all.
- **Location**. A fair ranges of countries and regions were represented in the data on this area. Study countries were the U.S. (six), Australia (four), the Netherlands (five), Indonesia (two), Germany (two), China (one), Israel (one), and South Korea (one).
- **Learning areas**. Disappointingly, the only subjects represented in our data were maths and science. This is despite our priority assessment looking to identify and prioritise studies from other areas to broaden the evidence-base. There were 17 studies of mathematical topics including algebra, geometry, subtraction, proportional reasoning, and fractions. Five studies were looking at science performance in physics (one), electrical circuits (two), biology problem solving (one), and chemistry problem solving (one).
- **Outcome measures**. The vast majority of the outcome measures were researcher designed (18/22). These typically consisted of problems aligned with the content of worked examples and other study materials used in the experiment. In some cases, there were tests designed to test transfer to similar knowledge or problem areas, usually in addition to bespoke content-aligned

tests. These tests ranged from a small number of items or problems to larger (for example, 46-item) tests but were generally in the vicinity of 6 to 12 items (for worked examples, usually complex problems). A small number of studies used regular curriculum assessments (two), standardised tests (one), or a mixture of standardised and research-developed assessments (one).

- **Design and delivery**. There was a mixture of formats for design and delivery. Our key concern (when assessing ecological validity) was who designed the learning materials and who delivered them. We estimate that 8 of 22 studies were both designed and delivered by the researchers. Another four used computer software designed by researchers and worked through independently by students. Two had lesson materials prepared by researchers and delivered by teachers using scripts or under supervision. Finally, eight studies were delivered by regular class teachers using a mixture of researcher and teacher-designed materials. McCann et al. (2019), for example, used lessons taught by the regular teacher as part of normal maths lessons; the teachers used their own lesson plans, adapted to align to each experimental condition.

High priority studies in this area

There were four studies for the worked examples strategy rated as having high strength and validity of evidence. We conducted in-depth analysis of these studies and have completed a full risk of bias assessment, summarised in the appendix.

Booth et al. (2015a). This study employed an experimental design to test the effect of the 'Algebra By Example' approach. The study included 380 eighth-grade students in 28 classes in five school districts in the U.S. The trial conditions were randomised at class level. The AlgebraByExample programme provides pairs of problems: a worked example accompanied by a 'your turn' item. Teachers can use the assigned problems as part of lessons in various ways including lesson 'warm-ups' and review and discussion prompts, both for independent and group work.[14] In the study, treatment students received a workbook containing interleaved worked examples and self-explanation prompts. There was a mixture of correct and incorrect worked examples. Control students were given the same problems to solve. The content was taught by the regular maths teacher throughout. They assessed outcomes using assessments of conceptual and procedural knowledge (66 items, researcher-developed) and ten items from standardised algebra curriculum tests.

Key findings. Students receiving the AlgebraByExample intervention received higher post-test scores for standardised test items and conceptual knowledge. The effect of the intervention was especially strong for conceptual post-test scores for students with low prior knowledge. Treatment students outscored control students by 7% on the items from the state standardised test. For students in the lower half of the performance distribution this increased to 10%. Treatment group gains were also seen on the assessments of conceptual and procedural knowledge of 5% and 4%, respectively. The risk of bias assessment identified some concerns with the randomisation process, missing outcome data, and selection of the reported results. This was mostly an issue of report and protocol. The latter, for example, requires studies to have pre-determined statistical analysis plans. Overall, this was graded as 'some concerns'.

Booth et al. (2015b). This study presented two experiments and like Booth et al. (2015a) focused on worked examples in algebra. Correct and incorrect worked examples were provided to eighth- and ninth-grade students in the U.S., embedded in a school algebra unit taught by the regular class teachers. Experiment 1 was smaller (n = 51, two high schools, three classes) than experiment 2 (n =

[14] https://www.serpinstitute.org/algebra-by-example

395, seven schools, 28 classes). Both were randomised at the individual level. The algebra pre- and post-tests were based on typical curriculum assessments.

Key findings. They found that worked examples led to better post-test scores than traditional problem solving (Experiment 1, but had no main effect in Experiment 2, the larger RCT). Students with low prior knowledge showed greater improvements when given worked examples than those with high prior knowledge (Experiment 2). The risk of bias assessment identified concerns with the reporting as the only potential issue. This study had 'low' risk of bias in all other categories.

Heemsoth et al. (2014). This study examined the effect of incorrect versus correct worked examples of fractions learning for 195 sixth-grade students in Germany in four high schools and nine classes. The trial randomised students at an individual level. Both groups studied a total of 21 worked examples, with the treatment group receiving incorrect examples and the control correct ones. The intervention took place across three weeks (11 lessons). Four introductory lessons were received by all (multiplying and dividing fractions) alternated with seven practice lessons in which students from the two conditions worked with different materials in the same room. All lessons were designed and led by researchers. The learning outcome was a research-developed test of fraction knowledge and another test of incorrect (or 'negative') strategy knowledge.

Key findings. Only advanced students benefited from studying incorrect examples; students with lower prior knowledge learned more from the correct examples. Students studying incorrect examples showed more negative knowledge than those studying correct examples. The risk of bias assessment identified concerns with the reporting as the only potential issue. This study had 'low' risk of bias in all other categories.

McLaren et al. (2015) tested the effect of learning decimals through erroneous examples using a web-based tutor. This was an RCT, assigned at the individual level, with 390 sixth-grade students in two middle schools in the U.S. In the treatment group, students learned about decimals through finding and fixing errors. This was compared with students learning through solving conventional decimals problems and explaining their solutions. The intervention used an interactive, web-based tutor and took place over two maths lessons in the school computer lab. All students were given correctness feedback by the programme throughout. The intervention effectiveness was assessed through a researcher-developed, 46-item decimal assessment test administered pre-test, immediately, and one week after the intervention. Students' preferences were also surveyed.

Key findings. McLaren et al. (2015) found no difference between groups for scores on the immediate assessment. The erroneous examples group, however, performed better on the delayed post-test (d = 0.33, 95% CI: 0.13, 0.53). In addition, the problem-solving group reported liking the intervention more than the erroneous group did (d = 0.21). The risk of bias assessment identified concerns with the reporting as the only potential issue. This study had 'low' risk of bias in all other categories.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section we focus on the study outcomes, summarised in the Table B4.2. Studies identified as high relevance and quality have been marked with an asterisk.

Table B4.2: Worked examples—summary of evidence

Study	Focus	Population	Finding
High Priority Studies			
Booth *et al.* (2015a)*	Effect of AlgebraBy-Example assignments on algebra test scores	*N* = 380 8th grade 5 school districts, 28 classes US	***Positive*** • Students receiving AlgebraByExample intervention received higher post-test scores for standardised test items and conceptual knowledge (standardised effect in random effect model = 0.06, SE = 0.03, $p < 0.05$). • Effect of intervention was especially strong for conceptual post-test scores for students with low prior knowledge
Booth *et al.* (2015b)*	Effect of correct and incorrect worked examples on algebra test scores	2 experiments Ex1 – *N* = 51, 2 high schools, 3 classes Ex2 – *N* = 395, 7 schools, 28 classes 8th/9th grade US	***Neutral*** • Worked examples led to better post-test scores than traditional problem-solving (Expt. 1), but had no significant main effect in Expt. 2 (larger RCT). For Expt.2 standardised effect in random effect model = 0.08, SE = 0.09, $p > 0.05$). • Low prior knowledge students showed stronger outcomes after using worked-example assignments than those with higher prior knowledge. • The authors discuss variation by topic area, arguing that targeting of misconceptions in topic areas matters.
Larger Studies (pupil n > 500) (Medium Priority)			
There were no larger studies at the medium priority level			
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
Bokosmaty *et al.* (2015)	Effect of worked example guidance type on geometry problem-solving	2 experiments *N* = 360/120 Years 8&9/7&10 1 school Australia	***Positive*** • Expt. 1: For all items, Step Guidance led to best performance (M=5.23-7.42, SD=0.81-0.93), theorem and step second-best (M=4.28-6.02, SD=0.93-1.12), unsupported problem-solving worst (M=2.30-4.52, SD=0.76-1.09). In both Year 8 and Year 9. Theorem plus guidance created higher cognitive load.
			Positive • Expt. 2: Expertise Reversal Effect: Year 10 students (experts) performed better when given Step Guidance Only, whereas Year 7 students (novices) performed better with Theorem *and* Step Guidance
Mevarech *et al.* (2003)	Effect of metacognitive training versus worked examples on mathematical reasoning	*N* =122 8th grade 5 classes Israel	*Worked example is the control. Included for info as comparison.* • Students in the metacognitive training group outperformed worked examples group at both immediate post-test (low achievers *ES* = .51; high achievers *ES* = .14), and delayed post-test (overall *ES* = .40) • No significant effect of prior knowledge
Mulder *et al.* (2014)	Effect of heuristic worked examples on inquiry-based learning in physics	*N* = 107 *M* age = 15.51 (*SD* = .42) School n not reported The Netherlands	***Neutral*** • No significant differences in post-test scores between experimental and control groups, though the worked example group did improve on some inquiry behaviours
Reed *et al.* (2013)	Effect of worked examples and Cognitive Tutor on constructing equations	*N* = 128 Age/grade not reported 3 high schools US	*No BAU control. Included for info as comparison.* • No difference between the 4 groups on any of the algebra test scores. • Cognitive tutor was not a worked example, but is not a BAU condition for a control.
Retnowati *et al.* (2010)^	Effects of collaborative learning and task complexity on mathematics performance	*N* = 101 7th grade 1 high school, 3 classes Indonesia	***Positive*** • Worked example approach produced greater test scores in both group and individual settings • Students reported a preference for worked examples across both conditions

Retnowati *et al*. (2017)^	Effects of collaborative learning and instructional format on mathematics performance	7th grade Expt 1: *N* = 182 1 high school, 6 classes Expt 2: *N* = 122 1 high school, 4 classes Indonesia	***Positive*** • Learning individually resulted in better performance for high- complexity tasks than learning collaboratively, but no difference for low-complexity tasks (Expt. 1) • Across all conditions, studying worked examples was superior to problem-solving (Expt. 2) • When studying worked examples, individual learning was superior to collaborative learning (Expt.2) • When problem-solving, collaborative learning was superior to individual learning (Expt.2)
Van Gog *et al*. (2011)	Effect of worked example format on learning electrical circuits	*N* = 103 *M* age = 16.22 (SD = .84) 2 high schools The Netherlands	***Positive*** • Both the examples only (M=4.8, SD=2.6) and example-then-problem (M=4.7, SD=2.8) pairs conditions were more effective than the problem-then-example (M=2.5, SD=2.3) and control conditions (M=2.7, SD=1.6)
Wong *et al*. (2019)	Effects of worked example type and self-explanation type on geometry test scores	*N* = 122 6th grade 1 middle school, 2 classes US	*No BAU control. Included for info as comparison.* • Process/product examples and focused/menu-based support compared. • No effects of self-explanation prompt type or worked example type on test scores • No differences in cognitive load between conditions
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Bentley *et al*. (2017)	Worked examples effect on proportional reasoning problems in mathematics	*2 experiments* *N* = 44/38 Ages 11-13years 1 high school, 1/3 class Australia	***Positive*** • Expt.1: WE group outperformed the control group on both post-tests, with effect sizes d = 0.89, 95 % CI = 0.27, 1.52) at T3 and d = 0.70, 95 & CI = 0.09, 1.30 at T4 ***Positive*** • Expt. 2: WE group outperformed the control group at both T3 (d = 0.89, 95 % CI = 0.23, 1.56), and T4 (d = 0.79, 95 % CI = 0.13, 1.45)
Kyun *et al*. (2009)	Effects of worked example presentations on algebraic problem solving	*N* = 97 Ages 12-14 1 middle school South Korea	***Positive*** • Using conceptual and procedural worked examples was the most effective condition in retention (d = .43) and transfer tests (d = .21)
Van Gog *et al*. (2012)	Effect of worked example type on learning electrical circuits	*N* = 82 *M* age = 16.10 (*SD* = .49) n of schools or classes not reported The Netherlands	*No BAU control. Included for info as comparison.* • Process-oriented examples conditions outperformed product-oriented examples (*d* = .46) • Then a second condition compared combinations. Process-product condition better than process-process condition (*d* = .50), but no better than product-product and product-process conditions • Expertise reversal effect: extra detail (i.e., process-oriented worked e.g.) may be useful initially but then may become redundant as training progresses
Van Loon-Hillen *et al*. (2012)	Effect of worked examples on subtraction performance	Quasi-experiment (assigned at class level) *N* = 45 4th grade 1 primary schools, 2 classes The Netherlands	***Neutral*** • No difference in post-test scores between the worked example and BAU conditions (d = 0.15, 95 % CI = -0.43, 0.74) • No difference in perceived cognitive load between groups (d = 0.01, 95 % CI = -0.06, 0.57)
Youssef-Shalala *et al*. (2014)	Effect of worked examples on	*N* = 40 9th grade	***Neutral***

– Expt. 3 only	geometry problem solving	high schools, ? classes Australia	• Results varied by tests (acquisition, similar test, transfer test, far-transfer test). No overall and consistent difference in test scores between worked examples and problem-solving groups (Expt. 3)

* High priority study identified for in-depth analysis; ^ = study included for more than one strategy.

We have separated the general worked example studies from those looking specifically at incorrect or incomplete worked examples, summarised in Table B4.3. We interpret *all* studies (B4.2 and B4.3) that test using worked examples as a strategy. Separating the studies may help us tease out any principles and provisos for the successful use of worked examples. The theory would suggest that the effect will be moderated by learner prior knowledge. The conditions below vary in terms of the conditions and whether subgroup analysis is conducted to identify this ability-example match. We have examined the effect of incorrect versus correct examples and whether the result is consistent with the theory that incomplete or incorrect worked examples will be increasingly beneficial as learners develop knowledge in the problem area. We discuss the results further below.

Table B4.3: Incorrect or erroneous worked examples—summary of evidence

Study	Focus	Population	Finding
High Priority Studies			
*Heemsoth *et al.* (2014)	Effect of incorrect examples on learning fractions	*N* = 195 6th grade 4 high schools, 9 classes Germany	***Positive (for advanced students)*** • Fraction knowledge: Only advanced students benefited from studying incorrect examples; students with lower prior knowledge learned more from correct examples (d = 0.68, 95 % CI = 0.06, 1.29) • Negative knowledge: Students studying incorrect examples showed more negative knowledge than those studying correct examples (d = 0.22, 95 % CI = -0.06, 0.50)
*McLaren *et al.* (2015)	Effects of learning with erroneous examples on learning decimals with a web-based tutor	*N* = 390 6th grade 2 middle schools US	***Positive*** • No difference on immediate test scores between groups (d = 0.03, 95 % CI = -0.17, 0.23) • Erroneous examples group performed better on delayed post-test (*d* = .33, 95 % CI = 0.13, 0.53) • Problem-solving group reported liking intervention more than erroneous group did (*d* = .21)
Larger Studies (pupil n > 500) (Medium Priority)			
There were no larger studies at the medium priority level			
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
Grosse *et al.* (2018)	Effects of copying correct and incorrect solutions on mathematical problem solving	*N* = 139 7th grade 1 high school Germany	***Negative (but no ability sub-group analysis)*** • Correct examples fostered greater performance in the learning and test phases (d = 0.91, 95 % CI = 0.56, 1.26) • While learning, presenting examples and problems simultaneously increased scores, but in the final test, simultaneous presentation during learning led to lower scores.
Yang *et al.* (2016)^	Effects of collaborative learning and erroneous examples on subtraction knowledge	*N* = 109 3rd grade 1 elementary school, 2 classes China	***Neutral (but no ability sub-group analysis)*** • No main effects of example type or social context on 3-digit subtraction knowledge • Interaction effect: students learning individually using erroneous examples showed improvements in 3-digit subtraction at immediate and delayed post-tests • Students learning collaboratively better able to apply knowledge to advanced (4-digit) subtraction than individual learners, but only when learning from correct worked examples
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Baars *et al.* (2013)	Effect of partially worked-out examples on biology problem-solving and students' judgements of learning	*N* = 66 Year 3 (*M* age = 14.61 years) 1 high school, 3 classes The Netherlands	***Neutral*** • No differences in test performance between complete and incomplete example conditions • Accuracy of judgements of learning were less accurate when studying incomplete examples (underestimated future performance – students were less confident)
McCann *et al.* (2019)	Effects of generating incorrect examples on algebra tests	*N* = 99 9th grade 1 high school, 4 classes US	***Neutral (no ability breakdown)*** • No significant differences in test outcomes between conditions (d = -0.24, 95 % CI = -0.81, 0.31)
Ngu *et al.* (2002)	Effects of text editing on chemistry problem solving	2 experiments *N* = 23 11th/10th grade 1 high school, 1 class Australia	***Neutral / mixed*** • Text editing group outperformed conventional problem-solving group for dilation (d = 0.17, 95 % CI = -0.07, 1.01) and molarity problem solving (d = 1.12, 95 % CI = 0.22, 2.02) (Expt.1) • Conventional problem-solving group outperformed text-editing group for stoichiometry problem solving (d = 1, 95 % CI = 0.13, 1.86) (Expt. 2)

* High priority study identified for in-depth analysis; ^ = study included for more than one strategy.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in the table below.

Table B4.4: Worked examples (including correct worked examples and incomplete or erroneous worked examples)—quality of evidence assessment (based on the GRADE approach)

Strategy	Worked examples
Number of studies	There are 22 studies in this area of which four were rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	Twenty-one studies are randomised experiments; one was a quasi-experiment (Van Loon-Hillen et al., 2012).
Risk of bias	Our risk of bias assessments on the high-quality papers identified some concerns for one of the papers for missing data and randomisation. All papers had some concerns with reporting, particularly the lack of (or non-reporting of) pre-determined statistical analysis plans. However, we do not judge this to have affected the results. We judge therefore, there to be at least three strong studies in this area.
Inconsistency	**Result consistency**. There was some inconsistency in the results, but mostly with a mixture of neutral and positive outcomes. Neutral outcomes were more likely for smaller studies and studies specifically of incorrect or incomplete worked examples.
Indirectness	**Practice and population heterogeneity**. The studies in the area had an unfortunate combination of being overly homogenous in terms of population (mostly secondary age) and subject (maths and science), while highly varied in terms of the strategies employed. While the instructional approaches varied considerably, we judged there to sufficient adherence to test the general concepts of cognitive load theory in the classroom. **Measure and outcome heterogeneity**. The vast majority of studies used researcher-designed bespoke tests aligned to the problems or tasks practiced in the intervention. **Design and delivery**. There were a mixture of interventions delivered by researchers and teachers. Given the focus on problem solving, many studies had the learners working independently on set tasks.
Imprecision	**Group sizes**. There were a mixture of small and small-moderate scale studies. There were no large trials in this area. As a result of this and heterogeneity—and a large spread of results—we judge these estimates to be highly imprecise. The studies reporting effect sizes that we judge to be the most precise are: - *Heemsoth et al. (2014): d = 0.68 (95% CI: 1.29, 0.06); - *McLaren et al. (2015): d = 0.33 (95% CI: 0.13, 0.53); - Booth et al. (2015a)*: standardised effect in random effect model = 0.06, SE = 0.03, $p < 0.05$; and - Booth et al. (2015b)*: standardised effect in random effect model = 0.08, SE = 0.09, $p > 0.05$.
Publication bias	There was no evidence of publication bias in these results.
Other considerations	Here we have combined results for correct and incorrect/erroneous worked examples. Results for the latter were more inconsistent. One issue with this is that the theory would suggest that incorrect or erroneous worked examples would be suitable for more advanced learners, but results did not consider this in several cases.
Overall confidence	**Moderate (+++)** We are moderately confident in the effect estimate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.
Confidence reasons	This result has been graded as moderate certainty, but with the important proviso that this applies only to secondary-age pupils in maths and science. There was some inconsistency in outcomes, but this was plausible given that these were concentrated in the incorrect and erroneous examples group and given the theory. Between the two groups, there was a sufficient weight of evidence to support this strategy. Limitations included: - studies focused on secondary science and maths;

	- the vast majority of studies used researcher-designed bespoke tests aligned to the problems or tasks practiced in the intervention; - there were a mixture of interventions delivered by researchers and teachers; - there were no large trials in this area; and - reporting was poor, making effect size estimates more challenging, and there was inconsistency in the outcomes that were reported.

Summary of findings for this strategy

Main finding. The evidence for all worked examples (including the incorrect and incomplete examples), suggests a small to moderate positive effect of using worked examples for secondary-age students in maths and science, compared to conventional problem-solving techniques.

Estimated impact. The evidence was, to some extent, mixed: there were several neutral results and some with a negative effect. High priority studies suggest standardised effect sizes ranging from 0.06 to 0.68. The lack of large studies limits confidence in specific effect sizes.

Confidence in impact estimate. Our assessment is that we can be moderately confident that the effect of worked examples for secondary science and maths is positive. Results were, however, inconsistent and there were too few high-precision studies to reach confident effect size estimates. However, the results were, with only one exception, neutral or positive. The exception—and several neutral studies—had convincing explanations for non-positive results. There was no evidence of publication bias. As a result, we have moderate confidence that the effect is positive but cannot estimate a range for the magnitude of the effect.

Heterogeneity. Many of the neutral results were concentrated in the 'incorrect and incomplete worked examples' studies where there appeared to be issues matching learners with the right level of support. Many of those studies did not provide a breakdown of students' abilities so we cannot make a confident judgement about incorrect and incomplete worked examples from the limited evidence we have. We return to discuss this point in the overall conclusion.

Strategy 6: Scaffolds, guidance, and schema-based instruction

Concise definition

Scaffolding is a general teaching strategy and concept where educators provide learners with guidance or other forms of support such as prompts, cues, instructions, information organisation, or reference materials. We focus specifically on scaffolding strategies designed to optimise cognitive load.

Full definition and description

Scaffolding is a general teaching strategy and concept where educators provide learners with guidance or other forms of support such as prompts, cues, instructions, information organisation, or reference materials. We focus specifically on scaffolding strategies designed to optimise cognitive load, other than worked examples (analysed separately under Strategy 5). These all support students to complete a complex problem or task and draw attention to the core schematic structure of the process or knowledge content of the task.

Selected examples

Examples of this strategy from our database.

- In De Corte, Verschaffel and Van De Ven (2001), students were supported with the process of reading text using four strategies: activating prior knowledge, clarification of difficult words, schematic representations of text (grids with key words and structure to support students to organise text into themes, sequences, and comparisons), and in formulating the main idea.
- Wijekumar et al. (2014) used an intelligent tutoring system to structure the reading of the text. There were instructions, prompts, words highlighted in the text, and grids for organising and recording key ideas.
- Becker et al. (2020) experimented with different ways of representing a common object in science. Their physics experiments were supported with graphic calculators and tablets with provided diagrams, strobe pictures, tables, and formulas. The authors tested whether the enhanced environments would lead to a reduction in extraneous cognitive load and better conceptual understanding.
- Fuchs et al. (2004) provided schematic prompts for students to use during problem solving that supported them to identify the common, transferrable features of the problem and support them to identify this schematic structure in ostensibly novel problems of the same type. Worked example (Strategy 6) was part of a more general approach including teachers explaining the steps, posters listing the steps, examples and discussion of transfer of the problem-solving schema across problems, and the use of an answer key.
- Richey and Nokes-Malach (2013) tested the provision (versus the non-provision) of stepwise instructional explanations while students solved problems. Control students received more sparse worked examples with solution steps and short process instructions—for example, 'write Ohm's law', 'this is what we must solve for', 'current is already isolated in this equation'. Treatment students received explanations at each step—for example, 'electric current is …' and 'an increase in resistance (R) leads to a …'.
- Olina et al. (2006) provided with a cue 'in the form of a comma rule that should be applied in each sentence' within a text editing task.
- Oksa, Kalyuga and Chandler (2010) provided explanatory notes with a Shakespearean text (and found an expert reversal effect, see Discussion section).
- Pawley et al. (2005) provided students with instructions to check aspects of the problem.
- Li and Liu (2007) explored whether a database tool could share sixth-graders (age 11 to 12) cognitive load in a problem-based learning environment.

Evidence for this approach

There were 16 studies for scaffolds, guidance, and schema-based instruction. As noted at the beginning of the section, there is a separate section on schema development. Here we have included a small number of studies self-describing as schema-based instruction where there was a significant component of using schematic diagrams, prompts, or methods to reduce cognitive load in a problem-solving space. The main groups within this are (a) providing targeted explanations to support learning, (b) providing schemas and structures to support students to manage tasks, and (c) providing supports that manage information during the activity. Of the 16 studies in this area, two were graded as high relevance and quality. Full details of all medium and high studies are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed for this strategy are characterised as follows:

- **Pupil age and characteristics**. The age range of students varied from third grade (age 8, Year 4) to tenth grade (age 15 to 16, Year 11). There was a good spread across this range with approximately two studies per year-group. This data, therefore, covers the second half of primary and the full secondary age range.
- **Location**. The majority of studies (ten) were from the U.S. There were also three studies in Germany, two in Australia, and one in Belgium.
- **Learning areas**. There was a range of learning areas represented by the studies. There were five studies with maths as a learning outcome, five with reading comprehension, one grammatical knowledge, one history, and four in science.
- **Outcome measures**. Most outcome measures were researcher-designed tests based on the specific target content from the intervention (11/16). Several studies used tests developed between teachers and researchers or a combination of standardised and bespoke tests (four). One used a standardised reading comprehension test (De Corte et al., 2001).
- **Design and delivery**. About half of the studies (seven) were taught by the learners' regular teacher. Most of these involved research-designed lessons and, in several research-designed scripted lesson plans, with lessons supervised by research assistants. The other sessions (eight, apart from one which was not reported) were delivered via computer software or a researcher-designed booklet. Teachers supervised the pupils as they completed the set work.

High priority studies in this area

There were two studies in this category that were rated as having high strength and validity of evidence. We conducted in-depth analysis of these studies and have completed a full risk of bias assessment, summarised in the relevant appendix. In summary, these studies were as follows.

Becker et al. (2020). This study examined the impact of providing tablets to students to support experimental learning in physics. The tablets provided and recorded measurement data to lower extraneous cognitive load. The control group was a 'strong' control condition. Students were provided with a graphing calculator that provided some diagrammatic and tabular output, so this study tested the effect of the tablet support (for example, overlays, changing representation, and dynamic linking) over and above the control support. The study employed a cluster randomised design with 286 students of mean age 15.6 at 11 secondary schools in Germany. After an introductory lesson in the technology and a pre-test, students conducted experiments using an experimenter-designed protocol booklet. The post-test was carried out immediately after learning. Learning was measured using a multiple-choice test consisting of adapted items from validated test instruments and researcher-designed and validated items.

Key findings. Becker et al. (2020) found a very small improvement on one of three learning measures ($\eta 2 = 0.015$) and no significant difference for two of the three. They also found that the approach led to a significant reduction of extraneous cognitive load, as intended. Our risk of bias analysis identified some concerns with the randomisation process and selection of reported results (specifically lacking a pre-planning analysis and reporting plan). All other areas (deviations from the intervention, missing data, and outcome measurement) were rated as low risk.

Wijekumar et al. (2014). The second study rated as high in the area, and one of the few large-scale studies in our entire database was Wijekumar et al. (2014). There were three studies (2012, 2014, and 2017) from Wijekumar and colleagues, all focusing on the effect of an intelligent tutoring system, at scale, on reading comprehension and recall. We have selected only one for in-depth analysis and risk of bias assessment; these were all rated within our screening tool as 'high' ecological validity and a 'medium' test of a cognitive science principle. All three studies are included in this strategy area.

Wijekumar et al. (2014) examined the effect of an intelligent tutoring system on reading comprehension of expository texts. This was a multi-site cluster RCT, randomised at class level. It involved 2,645 fifth-grade elementary school students from 45 schools and 128 classes in the U.S. Treatment group students used a web-based intelligent tutoring system (ITSS) for 30 to 45 minutes each week as a partial substitute for the traditional language arts curriculum. The ITSS provided extensive guidance to support their reading of the texts. We note that cognitive science strategies and feedback were built into the system, so we cannot rule out that these were the operative factors within the approach. Nonetheless, we felt that guiding and scaffolding the process of text analysis and comprehension was a key aspect of the approach and so locate the study in this section. Specifically, the authors provided the structuring provided by the system as follows:

> *The following steps are built into ITSS using different formats:*
>
> *1. Identify the overall top-level structure of expository text (for example, Comparison, Problem and Solution, Cause and Effect, Sequence, and Description) by identifying signalling words (Meyer, 1975) used in text to explicitly cue these structures (such as 'in contrast,' 'on the other hand,' and 'different' for the comparison structure).*
>
> *2. Write the main idea using patterns for each of the different text structures. For example, for the comparison structure the pattern is: _____ and _____ (two or more ideas) were compared on _____, _____ , and _____ (X number of issues compared).*
>
> *3. Organize reading comprehension and recall by using the structure and main idea.*
>
> *(Wijekumar et al., 2014, p.337)*

Students studied independently, self-paced, with sessions supervised by teachers who received training on the software. Control students did not receive this intervention and received regular teaching of the curriculum without the intervention. The intervention lasted six to seven months and assessments were made using a standardised test of reading comprehension, the Gray Silent Reading Test (GSRT), alongside researcher-developed tests of reading comprehension.

Key findings. They found that students in the ITSS condition scored higher on standardised tests (an effect size of 0.20) and four researcher-developed reading comprehension tests (effect size 0.15–0.53) than control students. Similar results were obtained in the 2012 and 2017 studies. The risk of bias assessment identified concerns with the reporting as the only potential issue. This study had 'low' risk of bias in all other categories.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section, we focus on the study outcomes, summarised in Table B4.5. Studies identified as high relevance and quality have been marked with an asterisk. The results are provided in full in the appendix.

Table B4.5: Scaffolds, guidance, and schema-based instruction—summary of evidence

Study	Focus	Population	Finding
High Priority Studies			
*Becker *et al.* (2020)	Effect of tablets providing measurement data to support experimental learning in physics.	N = 286 M age = 15.6 11 secondary schools Germany	***Neutral*** • Very small improvement on 1 of 3 learning measures (η2 = 0.015). No significant difference for 2 of 3. • The approach led to a significant reduction of extraneous cognitive load.
*Wijekumar *et al.* (2014)	Effect of intelligent tutoring guidance on reading comprehension of expository texts	*N* = 2,645 5th grade 45 elementary schools; 128 classes US	***Positive*** • Students in ITSS condition scored higher on standardised tests (ES = .20), on 4 researcher-developed text comparison tests (ES = .42, .53, .32, .26) and on 2 problem and solution test tasks (ES = .20, .15) than control students. ES were adjusted (in HLM), standardised differences. All ITSS coefficients were p<0.01. (See previous for description of ITSS)
Larger Studies (pupil n > 500) (Medium Priority)			
Wijekumar *et al.* (2012)	Effect of intelligent tutoring guidance on nonfiction reading comprehension	*N* = 2,643 4th grade 24 elementary schools; 131 classes US	***Positive*** • Students in ITSS condition scored higher on standardised tests *(d = 0.32,* 95 % CI = -0.02, 0.67) and on 4 researcher-developed reading comprehension tests (d = 0.47, 0.43, 0.32 & 0.47) than control students. • NB. "Web-based intelligent tutors … [provide] consistent high-quality modelling, practice tasks, built-in assessments, and strong and customized scaffolding and feedback to the learners." (p.8)
Wijekumar *et al.* (2017)	Effect of intelligent tutoring on recall of expository texts	*N* = 4,001 4th & 5th grade 45 elementary schools; 259 classes, US	***Positive*** • ITSS had a positive (but not always statistically significant) effect in improving both Grade 4 and Grade 5 organised memory structures, and improving reading comprehension Results reported as odds ratios: odds of treatment being low vs middle performance = 0.48 to 0.99 [CI range: 0.39 to 1.37]; odds of being high vs. middle = 1.20 to 2.43 [CI range: 0.83 to 3.10]. (See previous for description of ITSS).
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
De Corte *et al.* (2001)	Effect of schema-based text comprehension strategies on reading comprehension	*N* = 228 5th grade 12 classes Belgium	***Positive*** • Intervention group scored significantly higher on Reading Strategies test (correct application of reading strategies) • Intervention group better able to apply the reading comprehension strategies to different contexts (I.e., higher transfer test scores)
Fuchs *et al.* (2004)	Effects of schema-based transfer instruction on real-life mathematical problem-solving	*N* = 351 3rd grade 7 elementary schools, 24 classes US	***Positive*** • In Transfer Tests 1: SBTI (pre-post d = 3.69) and expanded-SBTI (ES = 3.72) outperformed control group • Transfer Test 2: SBTI (d = 1.95) and expanded-SBTI (d = 2.10) outperformed control group • Real-life problem solving: for Transfer Tests 3 (d = 2.71), and 4 (d = 1.91), expanded-SBTI group outperformed the other 2 groups
Fuchs *et al.* (2006)	Effects of schema-based instruction type on real-life mathematical problem solving	*N* = 455 3rd grade 7 elementary schools, 30 classes US	***Positive*** • For most outcomes, both SBI groups outperformed control group, but comparably to each other. • On most 'real-life' problem-solving question, SBI-RL group outperformed both SBI and control groups.

Hawlitschek *et al*. (2017)	Effects of a motivational prompt on history learning in an educational game	*N* = 150 *M* age = 15.03 (SD = .69) 3 high schools Germany	***Negative effect for guidance for already intrinsically motivating learning game*** • No significant effects of instruction type on basic recall (d = 0.03, 95 % CI = -0.29,0.35), but groups with explicit learning instruction had significantly lower transfer knowledge scores (d = -0.34, 95 % CI = 0.66, -0.19) • Explicit instructions led to higher extraneous cognitive load with explicit instruction (though did not mediate effect on transfer scores)
Jitendra *et al*. (2009)	Effect of schema-based instruction on mathematical problem-solving	*N* = 148 7th grade 1 school, 8 classes US	***Mixed*** • SBI classes outperformed students in control classes on problem-solving measure at post-test (d = 0.45) and delayed post-test (d = 0.56) • No differences on standardised maths test
Olina *et al*. (2006)	Effect of cues and presentation sequence on grammatical knowledge	*N* = 209 9th grade 1 high school, 13 classes US	***Neutral*** • No significant main effects of problem type (cued vs. conventional) or presentation sequence • For high-achieving students only, group with blocked-order presentation (*M* = 18.44) outperformed random-order group (*M* = 17.03) on achievement test
Pawley *et al*. (2005)	Effects of explicit instructions to prompt and support answer checking and prior knowledge on equation formation	Expt.1: *N* = 156 Expt. 2: *N* = 153 Grades 8 & 9 1 high school Australia	***Positive*** • At higher knowledge levels, no-checking group outperformed checking group (partial n2 = 0.07), but no difference in groups at lower knowledge levels (Expt.1) • Checking hindered performance at higher knowledge level, while giving an advantage at lower knowledge levels (expertise reversal effect) (Expt. 2) • Checking imposed greater cognitive load in higher knowledge group (Expt.2)
Salden *et al*. 2009)	Effects of tutored problem solving vs. fixed faded worked examples on mathematics performance	Expt.1: *N* = 57 9th & 10th grade Germany Expt. 2: *N* = 51 9th grade 1 high school, 3 classes, US	*Not BAU as control condition, provided for comparison* • Adaptive-fading condition produced highest post-test scores in controlled setting (Expt. 1) • Positive effect of adaptive fading compared to fixed fading (d = .74) when in a more ecologically valid school setting (Expt. 2). However, high attrition on small sample so no statistically significant difference in scores between groups.
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Li *et al*. (2007)	Effect of using databases on problem-based learning in science	*N* = 98 6th grade 1 middle school, 6 classes US	***Positive*** • Computer database (as instructional aid) groups scored higher on achievement test than both control groups (d = 0.30) • Computer database groups reported lower cognitive load
Oksa *et al*. (2010)	Effect of explanatory notes on reading comprehension and cognitive load (Expts. 1 and 3 only)	Expt.1: *N* = 20, Year 10 Expt. 3: *N* = 20, Year 10 1 high school, 1 class Australia	***Positive*** • *Othello:* The explanatory notes group outperformed the conventional text group for both microstructure test items (d = 1.28, 95 % CI = 0.31, 2.24) and macrostructure test items (d = 1.96, 95 % CI = 0.90, 3.03) (Expt.1) • *Romeo and Juliet:* The explanatory notes group outperformed the conventional text group for both microstructure test items (d = 1.66) and macrostructure test items (d = 2.95) (Expt.2) • Explanatory notes = lower perceived learning difficulty (I.e. cog load)
Richey *et al*. (2013): Expt. 1 only	Effect of adding explanations to worked examples on physics problem solving	*N* = 80 6th & 7th grade 1 high school; 4 classes US	***Neutral*** • No significant differences between groups on any learning outcomes (though marginally positive effects of withholding explanations on conceptual understanding) (d = 0.73, 95 % CI = 0.17, 1.29)

Roelle *et al*. (2015)	(1) The effect of focused processing prompts vs. general instructions and (2) reduced explanations with prompts vs. full explanations on science knowledge.	N = 80, 77 Age 12-15 Germany	***Positive*** • Reduced explanation and focused processing prompts performed best (M=11.71, SD=3.9), then complete explanations and focused prompts (9.63, 2.9), reduced explanations and general instructions (7.80, 3.5) and finally complete explanations and general instructions (7.41, 3.7). • Focussed processing prompts lowered extraneous load. Not reduced explanations. Learners with lower knowledge benefitted most from reductions and vice versa for higher.

* High priority study identified for in-depth analysis.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in Table B4.6.

Table B4.6: Scaffolds, guidance, and schema-based instruction—quality of evidence assessment (based on the GRADE approach)

Strategy	The provision of scaffolds, guidance, or schema-based support to solve problems or learning in complex tasks
Number of studies	There are 16 studies in this area of which two were rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	All studies are randomised experiments.
Risk of bias	Our risk of bias assessments on the high-quality papers identified some concerns with the randomisation on one study and with the reporting (analysis pre-planning) of both. We judge therefore, there to be at least one strong study in this area: Wijekumar et al., 2014.
Inconsistency	**Result consistency**. Effect sizes, where reported, are highly variable. The majority of studies gave a positive result, a significant minority neutral or mixed, and one negative, which—as discussed above—we feel was not wholly representative of the area.
Indirectness	**Practice heterogeneity**. Before the analysis, we determined all to be linked conceptually—that is, as strategies to support the learning of complex material through supports designed to lower cognitive load (but not specifically focused on the provision of worked example, as per the previous area). As a result of a conceptual rather than practical definition of the strategy, there is particularly large variation in practice in the studies. The main groups within this involve (a) providing targeted explanations to support learning, (b) providing schemas and structure to support students to manage tasks, or (c) providing supports that manage information during the activity. **Population, measure, and outcome heterogeneity**. There was a good range of students from age 8 to 16. Most studies were from the U.S. Most studies were either maths, reading comprehension, or science, with a roughly three-way split between these. **Design and delivery**. About half of the studies (seven) were taught by the learners' regular teacher. Most of these involved research-designed lessons and in several research-designed scripted lesson plans, with lessons supervised by research assistants. While not ideal from an ecological validity perspective, this was a strength relative to other areas we have analysed giving greater support to the results' external validity—its real-world, scalable applicability.
Imprecision	**Group sizes**. There were numerous medium to large studies in this area and several smaller studies. High priority studies, or medium to large studies of medium eligibility providing effect size estimates include: - Wijekumar et al. (2012): d = 0.32 (95% CI: -0.02, 0.67); - *Becker et al. (2020): very small improvement on one of three learning measures (η2 = 0.015); no significant difference for two of three; and - Fuchs et al. (2004): real-life problem solving/Transfer Tests 3 (d = 2.71) and 4 (d = 1.91).

Publication bias	Larger studies tended to provide a smaller effect, suggesting that publication bias might be present in these results.
Other considerations	The negative result was a study with a different focus to the others: it provided a motivational rather than an instructional prompt for a learning game, which—as the authors concluded—was redundant as the game was intrinsically motivational. In retrospect, this study did not belong in this group, but we have retained it within this analysis for transparency purposes.
Overall confidence	**Moderate (+++)** We are moderately confident in the effect estimate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.
Confidence reasons	This was another area (as well as Strategy 5) where, while there were limitations in the evidence and a large range of impact estimates, the results were (with one exception) positive and neutral. We have moderate confidence that the effect is positive but cannot estimate the range in which it falls. Key considerations for this certainty rating were: - effect sizes, where reported, are highly variable (including one negative); - there was a good range of students from age 8 to 16, although no evidence for the youngest children (age seven and below); maths, science, and reading were all represented. (neutral); - about half of the studies (seven) were taught by the learners' regular teacher (positive, relative to other areas); and - this strategy encompasses a large range of specific strategies in a range of contexts; our confidence is for the general principle of leaner support to manage cognitive load, but not a specific activity or application to a specific learning objective.

Summary of findings for this strategy

Main finding. Overall, evidence suggests that well-targeted scaffolds, guidance, or schema-based supports are an effective approach to support students to solve problems or learn from complex tasks in Key Stage 2 to Key Stage 4 (age 8 to 16) in a range of subjects.

Estimated impact. There was large variation in the results and few precise estimates of impact. We cannot estimate the range in which the effect falls. There was one negative result, which in our view is atypical: we are confident that the effect is greater than zero.

Confidence in impact estimate. We have moderate confidence that the effect is positive but cannot estimate the range in which it falls. This judgement is based on the absence of negative studies and application of the strategy across subjects, ages (albeit not for the youngest children), and contexts.

Heterogeneity. Two of the neutral results were in physics. The negative result was a study with a different focus to the others: it provided a motivational rather than an instructional prompt for a learning game, which—as the authors concluded—was redundant as the game was intrinsically motivational. In retrospect, this study did not belong in this group, but we have retained it within this analysis for transparency purposes. We have not carried out further heterogeneity analysis given the limitations in the evidence.

Strategy 7: Collaborative problem solving

Concise definition

Problem solving involves activities in which learners work together to complete a problem or complex task. Collaborative problem solving is a general teaching strategy and concept. Here we focus on a subset of this specifically designed to optimise cognitive load.

Full definition and description

Problem solving involves activities in which learners work together to complete a problem or complex task. Collaborative problem solving is a general teaching strategy and concept. Here we focus on a subset of this specifically designed to optimise cognitive load. Through collaboration, learners can potentially share information, attend to different aspects of the task, and provide peer scaffolds (see Strategy 6) to optimise cognitive load.

Selected examples

Examples of this strategy from our database include:

- In Zambrano (2019) learners solved algebraic problems in groups of three; the study examined the role of group-work experience for the success of this.
- Retnowati, Ayres and Sweller (2016) examined the effect of learners working in groups of three or four on (a) worked examples and(b) problem solving (without worked examples) in mathematics. They found that group work was less effective for the former, more effective for the latter.
- Learners in a study by Kirschner, Paas and Kirschner (2009) were given complex biology (heredity) problems and worked in groups of three, or individually, to solve them supported by procedural instructions.

Evidence for this approach

There were nine studies focused on the optimisation or reduction of cognitive load through collaboration when problem solving. Of these, one was graded as high relevance and quality. Full details of all medium and high studies are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed for the collaborative problem-solving strategy are characterised as follows:

- **Pupil age and characteristics**. There were five studies of students age 14 to 16, three with students in seventh grade (age 12 to 13), and one with younger children in third grade (age 8 to 9).
- **Location**. A range of locations was represented in the data. There were two studies from the Netherlands, one from South Africa, three from Indonesia, two from China, one from Ecuador.
- **Learning areas**. The majority of studies were focused on learning in mathematics (six). There were two studies in science (biology) and one on ICT web design.
- **Outcome measures**. There was one study that made use of a standardised achievement test. All others used a test designed by researchers aligned to the targeted learning content, or the origin of the outcome measure was unclear.
- **Design and delivery**. Most interventions were overseen by the regular class teacher, although we note that this typically involved general (behavioural) facilitation of a problem-solving task provided by the researcher rather than any non-scripted teaching activity.

High priority studies in this area

Only one study in the collaborative problem-solving strategy category was rated as having high strength and validity of evidence. We conducted in-depth analysis of this study and have completed a full risk of bias assessment, summarised in the relevant appendix.

Kirschner et al. (2011). This study examined the effects of collaborative learning and instructional format on biology test scores. The study was a 2 x 2 randomised experiment (assigned at the individual level) involving 140 students with a mean age of 15 (SD = 0.96) from one high school in the Netherlands. This was a 2 x 2 design in which students were randomised into one of two instructional formats: traditional problem solving or worked examples, and one of two group configurations—individual or a group of three. After introducing the topic (heredity in biology, 15 minutes), students took part in three- to seven-minute learning tasks involving determining phenotypes and genotypes from biological traits (for example, eye colour, hair colour). The problem-solving group were given the solution and had to determine how it was reached. The worked example group, similarly, but were given a worked example. The learning was assessed using a researcher-developed genetics test. There was also a cognitive load measure.

Key findings. The results showed that problem solving (higher cognitive load) led to higher post-test scores than worked examples when studying in a group. When studying alone, worked examples (lower cognitive load) led to higher post-test scores than problem solving. The results, therefore, align with theoretical expectations—that problem solving is a higher cognitive load task than learning from worked examples but that collaborative settings can reduce the cognitive load associated with complex tasks. The risk of bias assessment identified concerns with the reporting as the only potential issue. This study had 'low' risk of bias in all other categories.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section, we focus on the study outcomes, summarised in Table B4.7. The study identified as high relevance and quality has been marked with an asterisk.

Table B4.7: Collaborative problem solving—summary of evidence

Study	Focus	Population	Finding
High Priority Studies			
*Kirschner *et al.* (2011)	Effects of collaborative learning and instructional format on biology test scores	*N* = 140 *M* age 14.98 years (*SD* = .96) 1 high school The Netherlands	***Positive*** • When studying in a group, problem-solving (high cog load) led to higher post-test scores than worked examples (d = 0.54, 95 % CI = 0.05, 1.02) • When studying alone, worked examples (low cog load) led to higher post-test scores than problem-solving (d = - 0.58, 95 % CI = -1.08, 0.08)
Larger Studies (pupil n > 500) (Medium Priority)			
Dhlamini *et al.* (2013)	Effects of collaborative learning on mathematics performance	*N* = 724 10th grade 9 high schools South Africa	***Positive*** • Learners in the group approach showed greater improvement in maths test scores than those in the BAU condition (*d* = 1.80, 95 % CI = 1.63, 1.98) • Exploratory analyses suggest this is due to decreased cognitive load.
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
Retnowati *et al.* (2010)^	Effects of collaborative learning and task complexity on mathematics performance	*N* = 101 7th grade 1 high school, 3 classes Indonesia	***Neutral (positive for worked examples – inc. above, but neutral for group setting)*** • Worked example approach produced greater test scores in both group and individual settings. • Students reported a preference for worked examples across both conditions.
Retnowati *et al.* (2017)^	Effects of collaborative learning and instructional format on mathematics performance	7th grade Expt 1: *N* = 182 1 high school, 6 classes Expt 2: *N* = 122 1 high school, 4 classes Indonesia	***Negative for high-complexity*** ***Positive for problem solving*** • Learning individually resulted in better performance for high-complexity tasks than learning collaboratively, but no difference for low-complexity tasks (Expt. 1) • Across all conditions, studying worked examples was superior to problem-solving (Expt. 2) • When studying worked examples, individual learning was superior to collaborative learning (Expt.2) • When problem-solving, collaborative learning was superior to individual learning (Expt.2)
Yang *et al.* (2016)^	Effects of collaborative learning and erroneous examples on subtraction knowledge	*N* = 109 3rd grade 1 elementary school, 2 classes China	***Negative (against theory, easier correct examples better in group, harder incorrect individual)*** • No main effects of example type or social context on 3-digit subtraction knowledge • Interaction effect: students learning individually using erroneous examples showed improvements in 3-digit subtraction at immediate and delayed post-tests • Students learning collaboratively better able to apply knowledge to advanced (4-digit) subtraction than individual learners, but only when learning from correct worked examples

Zambrano *et al.* (2019)	the effects of prior collaborative experience and density/distribution of information amongst collaborative learners in maths	N = 240 M age = 15.58 (SD=.84) 1 High school Ecuador	*Individual problem solving not control condition. Provided for comparison.* • For performance, experienced groups significantly outperformed (M = .45, SD =.22) inexperienced groups (M=.31, SD =.29). It also showed that groups with low information density (M=.42, SD =.25) outperformed high information density (M= .36, SD = .27). • Concerning mental effort, experienced groups (M=6.47, SD =2.51) reported more mental effort than inexperienced groups (M =5.53, SD =2.58).
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Kirschner *et al.* (2009)	The effects of individual versus group learning (in triads) on biology test performance.	N = 70 M age = 15.4 The Netherlands	***Neutral (positive but not stat sig.)*** • In the learning phase, groups had higher performance (94% compared to 70%) and lower perceived mental effort. • The learning condition test-type interaction 'approached significance' (p= .052), suggesting that participants who learned individually performed better on retention problems (95% to 84%), while participants who learned in a group performed better on transfer problems 54% to 47%).
Retnowati *et al.* (2018)	Effects of collaborative learning of worked examples and prerequisite knowledge on mathematics performance	N = 54 7th grade 1 high school, 2 classes Indonesia	***Positive (expected relationship between (in)complete knowledge and group.)*** • When learners have incomplete knowledge on a topic, collaborative learning superior to individual learning • When learners have complete knowledge, individual learning superior to collaborative learning • Above effects operate as a result of cognitive load reductions
Zhang *et al.* (2011)	The effects of two collaborative learning strategies (Open-ended and Task-based) with an individualized learning strategy on individual ICT learning in a computer-based environment	N = 94 9th Grade 1 secondary school 3 classes Macao	***Positive (for open-ended but not task based)*** Means/SDs for the groups were as follows on the webpage design measure: • Individual (M=67.8, SD=2.76) • Task-based (70.4, 2.93) • Open-ended (79.1, 2.72) • Author conclusion: "Overall the cognitive measures collected consistently concurred with a CLT explanation for the effects. In conclusion, we argue that a collaborative approach can be more effective on a complex computer-based task; however, the conditions of collaboration are important and moderate the impact of the strategy".

* High priority study identified for in-depth analysis; ^ = study included for more than one strategy.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in Table B4.8.

Table B4.8: Collaborative problem solving—quality of evidence assessment (based on the GRADE approach)

Strategy	Collaborative problem solving or worked example study
Number of studies	There are nine studies in this area of which one was rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	Seven studies are randomised experiments, two are quasi-experiments.
Risk of bias	Our risk of bias assessments on the high-quality paper identified one area of concern relating to the reporting and pre-planning of analysis. However, all other areas were low risk of bias. We judge, therefore, there to be one strong study in this area.
Inconsistency	**Result consistency**. Results were mixed. However, the evidence can be argued (*post-hoc*) to be consistent with the theoretical expectations.

Indirectness	**Practice heterogeneity**. As discussed above, this section looked at a mixture of worked examples and problem solving and different strategies to increase task demands—incomplete information, complexity, and erroneous examples. It is hard to reach general judgements in this context and subgrouping studies to maximise homogeneity would leave each area with very few studies to assess. **Population, measure, and outcome heterogeneity**. Student ages ranged from 8 to 16. There were several, and diverse, regions represented. Most of the studies were focused on mathematics learning (six). In addition, there were two science (biology) and one ICT. **Design and delivery**. Most interventions were overseen by the regular class teacher. However, we note that this typically involved general (behavioural) facilitation of a problem-solving task provided by the researcher rather than any non-scripted teaching activity.
Imprecision	**Group sizes**. Studies in this area were moderate to small . The largest study, Dhlamini et al. (2013; n = 724) was a quasi-experiment using a non-equivalent control group. High priority studies and large and medium-sized medium priority studies providing effect size estimates were: - *Kirschner et al. (2011): d = 0.54 (95% CI: 0.05, 1.02); and - Dhlamini et al. (2013): d = 1.80 (95% CI: 1.63, 1.98).
Publication bias	There are only a small number of studies in this group. There is a slight suggestion that smaller studies are more positive than medium and larger studies.
Other considerations	These results suggest that the conditions for collaborative problem solving are important. We discuss the indicative evidence for what these conditions might be below.
Overall confidence	**Low (++)** Our confidence in the effect estimate is limited: The true effect may be substantially different from our estimate.
Confidence reasons	The group of studies was rated as having low certainty for the following reasons: - its relatively small size; - the two negative and two neutral results; only one negative result could be explained with the strategy theory; and - there was considerable practice heterogeneity in this group, with it difficult to assess the comparability of learning conditions and problem-solving tasks.

Summary of findings for this strategy

Main finding. Overall, the evidence is supportive of the theory but with some negative results and complexity to note in relation to the conditions in which collaborative problem solving might be effective. The limited evidence suggests positive effects in maths and science of collaboration for traditional problem solving, that complex or erroneous worked examples are best for individual learners, and that worked examples with incomplete knowledge are superior for groups.

Estimated impact. The most reliable estimate of impact in this area is provided by the single high priority study, Kirschner et al. (2011), which reports an estimated moderate effect of d = 0.54 (95% CI: 0.05, 1.02). Given the small size of this group, this is indicative only.

Confidence in impact estimate. We have judged this area to have low certainty of evidence. We note that this summary is based on a small evidence-base with particular limitations in relation to pupil age (eight of nine studies look at learners aged 12 to 16) and subject (six maths, two science, and one ICT). We discuss these limitations further in the overall area conclusion below. This conclusion should be considered in light of this discussion.

Heterogeneity. Cognitive load theory suggests that collaborative learning will lower cognitive load and improve learning for problem solving or highly complex tasks (including incomplete or incorrect examples) via a 'sharing' of the task load. There are some interesting examples of studies that have successfully manipulated experimental conditions to test in which contexts collaborative learning

does, and does not, work. Several positive results (Dhlamini et al., 2013; Kirschner et al., 2011; Experiment 2 of Retnowati et al., 2017) found groups to be more effective at problem solving with students reporting reduced cognitive load. Similarly, Retnowati et al. (2018) found that learners working collaboratively could support each other's *incomplete* knowledge on a topic and performed better in group settings. However, students were best off working individually when they had *complete* knowledge for the task. One study (Retnowati et al., 2010) provided a neutral result with no clear difference between the individual and collaborative learners. Retnowati et al.'s (2017) Experiment 1 found collaborative problem solving less effective than lone working when using high complexity worked examples, which we judged to be against the expectations of the theory. Yang et al. (2016) found a negative effect for erroneous worked examples in group work relative to students working independently yet a positive effect of group work when it came to correct examples. Our expectation would have been that group work would have made the more demanding task (erroneous examples) relatively more manageable. This might relate to the complexity of group interaction in identifying and correcting (rather than reinforcing) errors or the interplay of cognitive load and prior attainment (and therefore appropriate task difficulty).

Management of cognitive load—overall evidence summary and conclusions

Summary of results

In this section, we have reviewed 91 studies focused on the management of cognitive load. We identified three strategies for which we potentially had sufficient evidence to assess effectiveness. Our results for these are summarised in Table B4.9.

Table B4.9: Managing cognitive load—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Worked examples	Twenty-two, of which four were graded as high priority.[1]	Small to moderate positive effect of using worked examples compared to conventional problem-solving techniques.	Results were entirely concentrated in maths and science and secondary-age students (11–18 years old).	Moderate (+++)
Scaffolds, guidance, and schema-based instruction	Sixteen, of which two were graded as high priority.[1]	Well-targeted scaffolds, guidance, or schema-based supports are an effective approach to support students to solve problems or learn from complex tasks.	There was a good range of students from age 8 to 16. Most studies were either maths, reading comprehension, or science, with a roughly three-way split between these.	Moderate (+++)
Collaborative problem solving with worked examples or scaffolds	Nine, of which one was graded as high priority.[1]	The evidence is supportive of the theory that collaborative learning will lower cognitive load and support learning during problem solving or complex tasks; although there were some negative results and complexity.	Student ages ranged from 8 to 16. Most of the studies were focused on mathematics learning (six). There were 2 science (biology) and 1 ICT.	Low (++)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Conclusions about strategies in this area

Managing cognitive load

Our headline conclusions in this area are:

- Cognitive load has high potential relevance across the U.K. education system and for all learners and subjects.
- Overall, the evidence is promising and indicates the value and importance of teachers seeking to optimise learners' cognitive load.
- There are numerous studies showing appreciable positive effects for strategies to manage cognitive load within the evidence we have. There are also appreciable numbers of neutral and negative results, suggesting complexity in the principles and challenges of making it work in practice.
- Much of the evidence we have is highly concentrated in specific age ranges and subject areas. Tests of worked examples have almost exclusively focused on secondary maths and science.
- Considering worked examples and other forms of scaffolding (for example, support and guidance for complex learning or problem-solving spaces) together suggests wider subject and age applicability (age 7 to 16) of the principle and provides greater confidence in the overall result. However, we note that this confidence is in the value of optimising cognitive load *per se*, rather than a specific strategy for doing so or for specific learner needs.
- Ecological validity was low for many studies, limiting our ability to generalise the findings to real educational settings confidently.

Worked examples

The evidence was largely in line with the overall theory but suggests that as learners develop knowledge, only partial supports are required. It can be challenging to consistently identify best practices. For novice learners, however, the evidence is clearer and supports the use of worked examples to manage cognitive load and support learning.

There are many studies in this area, but there are limitations in their robustness (vis-à-vis internal validity) and ecological validity. The other limitation with *worked examples* is that all 22 studies we reviewed were studies of mathematics (17) or science (5), and the majority of studies were for secondary-age students (20/22). Thus, while the results support the use of worked examples in preference to unguided problem solving in secondary maths and science, we must stress that the limitations in the present evidence-base prevent judgements of effectiveness beyond these subjects.

We also examined *incomplete and incorrect worked examples* within the overall worked examples section. The overall theory suggests that as learners start to develop knowledge in an area, incomplete and erroneous working examples can increase (desirable) difficulty and enhance learning. Our results are, again, broadly supportive of this principle (in secondary maths and science) but the results were less consistent than for worked examples as a whole. There appear to be issues matching learners with the right level of support. Moreover, many of these studies did not provide a breakdown of students' abilities and so we cannot make a confident judgement about whether student ability or their developing knowledge in the problem area is a key moderator of the effect as hypothesised for these studies.

At the outset of the managing cognitive load results section, we describe how the theory relates to the *optimisation rather than minimisation or maximisation of working memory load*. Incomplete or incorrect worked examples will tend to lessen learners' cognitive load compared with unguided

problem solving but produce a higher load when compared with complete worked examples. According to theory, whether this is optimal significantly depends on pupil prior knowledge in the problem area. The mixed results in the incorrect and incomplete worked examples section can therefore be interpreted as being in line with the overall theory. Still, the evidence is limited and suggests that it is difficult to make work in practice.

Scaffolds, guidance, or schema-based supports

The evidence suggests that *scaffolds, guidance, or schema-based supports* effectively support students to solve problems or learn from complex tasks. A wider range of pupil ages and subjects were represented in this data giving us greater confidence that the strategy is more widely applicable. However, the downside of this diversity was high heterogeneity in the learning aims, subjects, procedures, and assessments within this group of studies. The grouping of these studies was on a conceptual rather than practical basis. The practices were very different but we judged (prior to analysis) that all studies focused on learning complex material with supports designed to lower cognitive load (but not specifically focused on the provision of worked examples, as per the previous strategy). The main groups within this are (a) providing targeted explanations to support learning, (b) providing schemas and structures to support students to manage tasks, and (c) providing supports that manage information during the activity. Our overall 'moderate' confidence in our judgement that this is an effective general strategy comes with the caveat that we have specified the strategy at such a general level that it encompasses a huge range of practical strategies.

The other consideration is *how* the various supports used in this group of studies are conceptually and practically similar to those examined for worked examples. There were certainly many surface similarities, and it might be argued that some of the scaffolds we looked at in this section were the equivalent of worked examples—in particular, incomplete worked examples—but for a wider range of subjects. Subjects outside maths and science often have learning content that does not lend itself to specific and distinct (or algorithmic) problem-solving processes, for example, and so scaffolds, guidance, and schema-based support might be needed to manage cognitive load effectively. If this parallel is reasonable, we might look at the evidence in both areas collectively (note that both had an overall positive result with moderate confidence). Our categorisation of these studies was conducted before the analysis and separated out these two strategies. Future work, with greater attention to the specifics of strategies used, may wish to consider these collectively within a more granular taxonomy of the strategies and their contexts.

Collaborative problem solving

Finally, in relation to *problem solving*, our results suggest (a) positive effects of collaboration during traditional problem solving, (b) that complex or erroneous worked examples are best for individual learners, and (c) that worked examples with incomplete knowledge are superior for groups. However, we note that this summary is based on a small evidence-base with particular limitations in relation to pupil age and subject. Our confidence in this finding is low. Our judgement is that the complexities of task demands and the dynamics of group learning make clear principles about effective strategies more challenging. There is good evidence here that working collaboratively can lower cognitive load. Whether this optimises it for all learners, and the principles of how to do so, is a question that goes beyond the limited evidence we have and is a question we return to in the discussion and questions section.

Evidence-informed discussion and questions

Principles and moderating factors

How does one optimise working memory? In which teaching and learning activities and contexts should teachers seek to (a) reduce (extraneous or total) cognitive load and (b) increase (germane) load? Which situation is more typical?

As highlighted in the opening description of cognitive load, the central idea that underpins this group of strategies relates to understanding and managing working memory load. In the studies we reviewed, the 'mechanism' of working memory optimisation was present in all studies; in many cases, tests of working memory were built into the study. We have reported these alongside the findings. While the theory holds that optimisation can involve either increasing (germane) load or reducing load (especially extraneous load), in most studies we reviewed, the focus was on optimising working memory by reducing the burden on working memory in the context of high-demand tasks (that is, reducing the amount of information the learner needs to hold in their mind at any one time). The focus on reducing load is congruent with literature that demonstrates the role of working memory in goal-directed attention—that overloaded working memory increasing susceptibility to distraction (Lavie and Fockert, 2005). Put simply, as working memory is overwhelmed, the task becomes less manageable and the student becomes increasingly distractable and unable to attend to the task requirements.

While the central idea of optimising working memory (usually reducing the working memory burden in high-demand tasks) united all studies, we identified 93 studies in total in this area and were only able to group 45 into the three groups of broadly homogenous strategies. In this section, we report key findings from across the other remaining 48 studies along with wider evidence and discussion of the effectiveness principles for managing cognitive load. With too few studies in each area, we have not been able to conduct a systematic appraisal of the evidence. Nonetheless, we view the studies below as providing an important conceptual landscape for understanding the management of cognitive load. Future studies that test the principles and parameters of cognitive load management in ecologically valid classroom settings would help advance knowledge in this area.

This entire section comes with an important 'health warning': that its aim is an exploratory description of perspectives on theory and practice of cognitive load management. It is not a description of approaches and ideas that can be firmly linked to evidence from applied cognitive science (as above). As such, unless we explicitly refer to specific studies in our wider evidence-base (rated as medium priority but not grouped with the main strategy assessments), the reader can assume that we are relaying plausible (but often contestable) 'sense making' accounts of the science aimed at a practitioner audience.

Working memory (WM) training

Is it possible, feasible, and valuable to train working memory? Should schools be buying in or delivering working memory improvement programmes?

Our searches identified a group of 19 studies testing working memory training programmes. These fell outside of the main focus for this review, centred on the most prominent cognitive science strategies in use by teachers. Furthermore, because we did not search for working memory *training programmes* explicitly, our database does not contain an exhaustive list. However, these studies are of note for two reasons. First, the results from working memory training interventions shed light on the core principle that working memory is limited, and examines whether training working memory might be a fruitful complement to the strategies to managing cognitive load we have reviewed. Second, our systematic searches included dedicated terms for working memory and have located many studies in this area.[15] Researchers may wish to supplement our searches to identify specific working memory programmes and conduct a dedicated review of this area (our results suggest that there may be sufficient homogeneity in programmes and weight of evidence to consider meta-analysis).

We examined these 19 papers to identify those presenting the strongest causal evidence in terms of the use of a randomised research design, a large number of pupils and schools (and therefore statistical power), and absence of methodological issues such as attrition or other forms of bias. This assessment was not systematic or comprehensive and was designed to provide an indication of evidence in this area rather than a systematic assessment. An overview of the results of the five strongest studies of working memory training is provided in Table B4.10.

Table B4.10: Summary of selected studies of working memory training within our evidence-base

Study	Programme	Design	Findings
Dunning et al. (2013)	Cogmed Working Memory Training	Ninety-four children (M age 8y 5m, SD = 8m) were identified in screening as having low WM from 810 children attending nine schools in the North-East of England. Participating schools were randomly assigned to adaptive training, non-adaptive training, or no intervention conditions.	Adaptive training was associated with selective improvements in multiple untrained tests of working memory, with no evidence of changes in classroom analogues of activities that tax working memory, or any other cognitive assessments. Gains in verbal working memory were sustained one year after training.
Hitchcock and Westwell (2016)	Cogmed Working Memory Training	Primary school children (mean age = 12 years, N = 148) were cluster-randomised to complete active CWMT, a nonadaptive/placebo version of CWMT, or no training.	CWMT did not improve control of attention in the classroom, or regulation of social, emotional and behavioural difficulties.
Roberts et al. (2016)	Cogmed Working Memory Training	Population-based randomised controlled clinical trial of first graders from 44 schools in Melbourne, Australia, who underwent a verbal and visuospatial working memory screening. Of 1,723 children screened (mean [SD] age, 6.9 [0.4] years), 226 were randomised to each arm (452 total).	Of the four short-term and working memory outcomes, one outcome (visuospatial short-term memory) benefited the children at six months (effect size, 0.43, 95% CI: 0.25–0.62) and 12 months (effect size, 0.49, 95% CI: 0.28–0.70) but not at 24 months. There were no benefits to any other outcomes; in fact, the math scores of the children in the intervention arm were worse two years after the training.
Rode et al. (2014)	17-session, adaptive,	Grade 3 children received either a computerized working memory training for about 30 minutes per session (n =	Results indicated strong gains in the training task. However, effect sizes of training-specific transfer gains were very small and

[15] The search string in this area was: "working memory" OR "short-term memory" OR (load AND (Cognitive OR intrinsic OR extraneous OR germane))

	computerized WM training programme (unnamed)	156) or participated in regular classroom activities (n = 126).	not consistent across tasks. These results raise questions about the benefits of intensive working memory training programmes within a regular school context.
Wright et al., (2019)	Working Memory Plus intervention (WM+)	The evaluation was run as a randomised controlled trial with 1,475 pupils in Year 3 (aged 7–8 years) across 127 primary schools, randomised at school level to three groups: the Working Memory Intervention (42 schools), the Working Memory Plus (41 schools) intervention, and a control condition (44 schools). The control condition comprised a 'business as usual' approach where schools continued with normal classroom teaching and support for eligible pupils.	Children in both the WM and WM+ schools made the equivalent of three additional months' progress in maths, on average, compared to children in the business as usual control schools. These results have high security ratings. The evaluation found positive impacts on working memory and attention and behaviour in class for pupils receiving the interventions compared to children in comparison schools.

Information adapted from study abstracts or executive summaries.

With the exception of Wright et al. (2019), these studies suggest that working memory training improves performance on working memory tasks but that these effects do not tend to transfer to other outcomes. It is unclear whether the various positive results in Wright et al. stem from an improvement in WM, mnemonic and learning strategies, or content and skills from the exercises used to train WM. The neuroscience literature suggests that, while working memory is limited, it is only possible to improve capacity up to a certain point (Constantinidis and Klingberg, 2016). The positive result from Wright et al. (2019) combined with the basic science in this area suggests that there is value in further exploring the potential for WM interventions that are designed around or incorporate curriculum-relevant tasks (such as arithmetic) or produce effects that transfer to curriculum-relevant tasks.

Working memory, problem solving, and schema development

The problem with theory and wider evidence with respect to limited working memory (discussed at the outset of this section) is especially apparent in relation to problem solving. The evidence we have reviewed concerns strategies specifically designed to support students to navigate problems or complex tasks in line with cognitive load theory (see Sweller, 1988). This evidence broadly supports the theory that cognitive load is high in unguided problem-solving tasks, that this is detrimental to learning, and that using worked examples and other forms of scaffolding is effective. These instructional supports offer the dual benefit of reducing cognitive load while providing a more-expert schema for task completion and understanding the learning content.

The sources we consulted during the practice review connected several other ideas to these core claims. We are not able to take an evidence-based position on these based on the applied evidence we have and report these as hypotheses to explore rather than evidence-based claims.

Can the cognitive load of unguided problem solving for novices (also) be managed through developing generic or subject-specific (for example, maths) problem-solving skills?

One claim connected to this question is that domain or problem-independent knowledge or skills are limited in value,[16] with each (substantive) problem requiring domain or problem-specific prior knowledge. Therefore, the argument goes, teaching learners a general problem-solving strategy is not sufficient to manage and optimise learning from complex tasks. It is through having more established

[16] Many sources we consulted went further to say that there is no domain/problem-independent knowledge.

schemas that more advanced learners can complete and benefit from complex problems. Specifically, the ability of advanced learners to employ well-developed and connected knowledge (that is, schemas, see next section) in the complex space is what allows their long-term memory and working memory to be used in combination to complete the task. When established schemas are not present (in a novice) the learner will not successfully navigate the task and high cognitive load will prevent efficient or meaningful learning. As schemas develop, support becomes increasingly less necessary and less desirable.

Within this perspective, worked examples are an ideal solution in that they provide learners with a way of structuring and working through complex tasks, attending to each aspect in turn, without being overwhelmed. Furthermore, the worked example puts learners in a position gradually to familiarise themselves with and incorporate the schema and knowledge of the constituent material into long-term memory. One example of this account of problem solving, cognitive load, and worked examples is found in Kirschner and Hendrick (2020), who frames this as follows:

> *'The goal of instruction is not to have learners search for and discover information, but rather to give them specific support for guidance about how to cognitively manipulate information in ways that are consistent with a learning goal, and store the result in long-term memory. Approaches which achieve this are: modelling with and without explanations, worked/worked-out examples which are faded into partially worked examples and finally are faded into conventional tasks without support (see Van Merrienboer and Kirschner, 2018), process worksheets, and so forth.*
>
> *(Kirschner and Hendrick, 2020, p.170)*

One reflection we have about this view relates to the challenge—both definitional and practical—of determining (a) what information is relevant for a problem space and its pertinence, (b) what knowledge and skills are needed to 'manipulate information in ways that are consistent with a learning goal' in a specific problem space, (c) the relevance of knowledge of problem-solving strategies from related but not identical areas (such as the same subject area), and (d) what generic, domain-independent problem-solving skills are necessary (even if they are not sufficient). In other words, the connection between skills, and knowledge, and the import of their gradation in terms of specificity (versus generality) is not entirely clear from many accounts. It is easier to defend the position that completely unguided problem solving for novices is not effective, or that wholly generic problem-solving skills are not useful for problems requiring domain-specific knowledge; and it is even more defensible when dealing with defined problem areas with formulaic solutions. There is, however, a large continuum between unguided versus guided problem solving, complete novices with underdeveloped schemas versus students with advanced schemas, generic versus specific knowledge and skills, and defined versus undefined problem spaces and solutions. Moreover, with the suggestion that, as students develop schemas, the level of guidance should fall and problem complexity should increase, this grey area is one that teachers must navigate. Unfortunately, the vast majority of the applied cognitive science studies we reviewed in the main results focused on specific types of secondary maths and science problems and it has not proved possible to examine some of these subtleties. This all suggests that a key practical question for the application of cognitive load theory is ascertaining how much cognitive load a given learning task for a given subject area might induce, and how much information and element interactivity (versus guidance and support) is appropriate for a given learner with their current state of knowledge. Kirschner and Hendrick's account for practitioners (2020, p.169) states that cognitive load depends on (1) the number of novel learning elements and (2)

the level of interaction between them. These factors—which also appear in many studies we reviewed—appear to be a sensible starting point to explore this question in connection with realistic curriculum resources and activities.

Split-attention effects

What are the practical implications of split-attention effects? How serious are these and how should teachers minimise them?

A small number of studies' theoretical propositions (and accounts for practitioners in our wider evidence) examined various forms of extraneous load on working memory. One group of these related to split-attention effects. Studies such as Cerpa et al. (1996) and Purnell et al. (1991) examined the split-attention effect. Cerpa et al. (1996) compared learning through a computer programme, where students needed to split their attention between a manual and on-screen information, with having this information integrated in the programme on screen. They found that the integrated information group outperformed the split-attention group. Similarly, Purnell et al. (1991) looked at the split-attention effect in geography where students learned from maps and diagrams. Students often needed to split their attention between information in the diagram or map and the associated key or descriptors. They concluded that split attention can cause heavy cognitive load and impair learning. Other sources we have consulted suggest that splitting information over time (as well as space) can influence cognitive load. Breaking down a larger more complex task or topic is a common approach designed to lower cognitive load and allow pupils to learn component parts before bringing them together. The split-attention affect suggests the value of careful consideration about how information is grouped, learnt, and then subsequently integrated. Further research is needed to reach a sufficient weight of evidence to evaluate these propositions and their practical implications.

Anxiety and cognitive load

How substantial are the effects of emotion on cognitive load and learning? Are there practicable and effective strategies to support children with emotions that prevent cognitive overload becoming an issue?

There is a growing body of work exploring the impact of emotion on cognitive load during learning (for a detailed discussion, see Plass and Kalyuga, 2019). A small number of respondents mentioned stress and anxiety in relation to cognitive load in our questionnaire and interviews. In addition, several studies in our database explored the impact of anxiety, particularly maths anxiety on test performance. However, it is important to note that we did not systematically search for studies on anxiety and cognitive load and so our evidence-base for this is not exhaustive. Two studies, both rated as 'medium priority' in the review, specifically related anxiety to the taking-up of working memory resources and cognitive load.

- In a study of maths anxiety among primary school children, Mavilidi et al. (2014) found that stimulating students to look through the problems of a math test before the test resulted in less working memory resources being consumed by intrusive thoughts, and consequently, more resources were available for solving the maths problems.
- Allen and Vallée-Tourangeau (2014) explored the impact of anxiety in simple additions tests involving different lengths of additions and levels of interactivity (participants being able to touch the tokens, point to them, or manipulate them as they saw fit). They found that mathematics anxiety significantly predicted performance in the low-interactivity condition but not in the high-

interactivity condition. Based on this, they argue that working memory resources are augmented through interaction with the physical problem presentation, defusing the impact of anxiety on performance.

Such studies and our wider reading suggest that emotions, and particular anxiety, may be important for researchers and practitioners to bear in mind when considering cognitive load management. Going beyond an 'information processing' aspect of cognitive load to consider these emotional and relational (for example, the collaborative learning evidence above) aspects appears to be important for providing a complete account of the major factors in this area. Indeed, this point may be made of the prevailing professional account of cognitive science more generally. Saying that, we note that several professionally-focused accounts of the science and its implications for learning are already including emotion in their account. One example of this is the discussion of anxiety and strategies to overcome it in Agarwal and Bain (2019, Chapter 8).

More generally, there were several instances in our wider review of authors who—when summarising and presenting their own understanding of the science for practitioners—link emotions, motivation, or social aspects of learning to cognitive load. Didau and Rose (2016), for example, briefly discuss motivational aspects of challenges linked to cognitive load and expertise, as follows:

> *'A major difference, therefore, between the novice and the expert is that the former is more likely to become frustrated, and encounter cognitive overload, while the expert will remain interested. Unless highly motivated to succeed, novices need to experience some measure of success or they risk becoming demotivated as they encounter increasing difficulty. Conversely, experts are more likely to become ever more motivated by challenge.'*
>
> *(Didau and Rose, 2016, p.73)*

An early years teacher whom we interviewed also framed the issue of cognitive load in connection with some of the home-life challenges children face:

> *'The main differences for us between [a number of disadvantaged] children [and more advantaged peers] is how open to learning they are. So, the social and emotional aspects of their learning seems to take priority. We adapt to where these children are. Some come in, sort of in survival mode, some are thinking about what is going on at home. Some come through the door, this is a minority, not wanting to come to school because they are worried about mum and don't want to leave mum. Thinking of this in terms cognitive load, their minds are elsewhere, so it's what strategies that can kind of reduce that. So, we've got some good staff that kind of triage the children on entry. You can see by body language and then I think good teaching is good teaching.'*
>
> *(Interviewee 1)*

Variation in the practice or teaching and learning context

Pupil differences

Does cognitive load management differ in importance for different pupil groups, such as pupils with special educational needs or lower-attaining pupils? Are there differences in cognitive load management strategies and their applicability?

One thing we stressed in our main results was the limitation in terms of pupil ages represented within the evidence. All evidence was restricted to pupils aged 8 to 18; for worked examples, it was exclusively secondary—11 to 18 years old. This makes drawing conclusions about the applicability of cognitive load management and how it might be achieved for very young children (seven years old or under) highly uncertain. Popular, practice-facing guidance and the teacher questionnaire and interview respondents frequently discussed the value of managing cognitive load for different pupil groups, and in particular for students with lower prior attainment or specific educational needs. Chapter 4 of Harrington et al., (2020), for example, discusses specifically the impact of working memory and learning difficulties in the classroom and strategies for supporting students with working memory limitations. Several responses from our questionnaires and interviews related to cognitive load and pupils with additional needs, including:

- 'I think a lot of these strategies, because they allow us to break up, specifically with cognitive load, we have to be aware that students with SEN might become quicker overloaded because of their educational need … Children with SEND, we can still use the approach, just the chunks we still tailor and adapt. They are still doing the same content and following the same approach, but just adapted to their level. And that's one of the great things about these approaches, they can be adapted to any kind of pupil needs (Interviewee 11).'
- 'Planning tasks with cognitive load in mind is really effective for chunking and scaffolding work—particularly for classes with high SEN need and low reading ages. I am able to deliver challenging content (keeping expectations high) but in small sections and the pupils build the pieces up to like a jigsaw puzzle. This ensures they don't get overloaded, and if they do have difficulties it is very easy for me to spot the exact stage that is tripping them up allowing me to quickly respond to them.'
- 'SEN students benefit from reduction in cognitive load and often using pictures to support concepts, but as memory is so hard for many, the other strategies barely work—but it does depend on the SEN need, so has to be tailored to the individuals.'
- 'Managing cognitive load is most important to those who struggle more with memory. For the highest ability, they want the extra information and stretch—it inspires and enthuses them.'
- 'I think these are examples of quality first teaching which benefits all pupils, however awareness of cognitive load definitely supports those with poor working memory [and] processing issues.'
- 'Strategies to reduce cognitive load work particularly well for struggling learners.'

Gradually reducing support (fading) and learner generation

Is there a difference in support needs for pupils with lower and high prior knowledge in the problem area? What mechanisms might explain this?

In relation to sequencing (below)[17] and in many other areas, prior learning was raised as a key moderating factor for how tasks impact cognitive load. Several popular accounts frame the problem in terms of the distinction between 'experts' and 'novices'. Common ideas within these accounts relate to differences between how experts and novices tackle problems. The tenets of this popular view go as follows: expertise goes beyond knowledge since experts can often identify 'deeper' structures within a problem and are able to draw on their familiarity with the domain to identify its most pertinent features. Expertise also brings a degree of automaticity to tasks meaning that experts often negotiate a problem using steps or processes that would be unsuitable for novices. Novices, on the other hand, are more likely to attend to superficial features of a problem and do so under high cognitive load and therefore—as per the basic cognitive science—with far greater susceptibility to distraction (Lavie and Fockert, 2005). Didau and Rose (2016) provide the analogy of the difference between a wordsearch and a crossword for understanding differences in the level of processing. Both wordsearches and crosswords require processing of the word, but the former only at the level of identifying the word's opening letters, and the latter at the level of meaning (semantic) with cryptic crosswords requiring advanced knowledge of the focus word to solve.

We remind the reader that the current applied cognitive science evidence-base has not been sufficiently robust, extensive, or organised[18] for us to interrogate the evidence to assess the veracity of the above account as an explanation of the role of expertise in problem solving. What *was* apparent in the evidence we reviewed was that prior attainment was an important and significant factor, although the reasons for this were not clear. Studies in the main review—such as Oska (2010) and numerous studies conducting subgroup analysis to identify effects for high versus low prior knowledge learners—support the view that prior knowledge is a key moderating factor, so much so that effects can be reversed, from positive to negative, from low to high prior knowledge learners (the 'expert reversal effect'). Many studies explored the implications of this principle and how teachers can increase or reduce the level of support according to learner's prior knowledge and as the learner's knowledge develops. Some studies, such as Ardac and Unal (2008), focused on student differences—in this case, in student's ability to work with symbols—and in Schneider et al. (2019) on differences relating to student age. All of these studies suggest that prior knowledge is a relevant factor for cognitive load and its management, but the weight of evidence is insufficient to assess the specific mechanisms behind this or to identify any particular practical implications.

These studies support the view that prior knowledge is an important factor and that expert reversal effects can occur. Ostensibly, the implication of this is the need to gradually fade support as knowledge develops. This is based on the idea that (1) scaffolding of learning helps shift memory demands from extraneous cognitive load to germane/intrinsic cognitive load, which (2) leads to the development of schemas and storing of information in the long-term memory, which (3) means that learners require decreasing amounts of support (Kern and Crippen, 2017), and (4) so scaffolds and support should be reduced. Most popular accounts we consulted supported the principle of fading, with some exceptions such as Kirschner and Hendrick, (2020, p.170) who claim that 'for students with considerable prior

[17] In, for example, Song (2016) and Clark et al. (2005).

[18] That is, tightly organised around common methodologies and definitions to systematically identify and develop theory at an increasingly complex and granular level.

knowledge, strong support and guidance while learning is most often found to be equally effective to unguided and minimally guided approaches'. We review the small amount of evidence we located relevant to the following question.

Is there indicative evidence that gradually reducing support as learner knowledge develops ('fading') is beneficial?

Relevant medium priority studies in our review database were as follows:

- McNeill et al. (2006) explored whether continuous written instructional support or faded written instructional support would best prepare seventh-grade science students to write scientific explanations when the support was no longer there. They found that the faded group gave stronger explanations when they were no longer provided with support, giving support to this type of instruction.
- Kern and Crippen (2017) evaluated two scaffolding strategies for science instruction in secondary school: self-explanation prompts and faded worked examples. Cognitive load was the theoretical rationale for the study since self-explanation prompts and faded worked examples are successful in college-aged participants. However, Kern and Crippen did not find any advantage of these strategies compared to general prompts and therefore concluded that they might be less useful in pre-college populations.
- Finally, Salden et al. (2008) compared the effect of faded worked examples that either occurred in a fixed manner or were adapted to the individual students' understanding of the examples by an intelligent software tutor. They found that the adaptive fading option improved learning compared to the fixed fading and argued for matching the fading procedure to the changing knowledge level of individual students.

Faded examples were a common feature of teacher descriptions of cognitive science informed practice in our questionnaire (see further detail below). Several questionnaire respondents also discussed the challenges of fading scaffolds in their classroom. Three comments made (by different respondents) are as follows:

- 'Students tend to keep using the scaffolds we gave them early in the school (KS3) and continue using them for KS4 and even KS5 work, resulting in restricted, simplistic writing (especially for science investigations).'
- 'Knowing how much to scaffold, how much to break down—this is about assessing prior knowledge accurately and that can be difficult.'
- 'I'm struggling to remove scaffolds and get students to work independently.'

Several studies in the area of cognitive load also examined the idea of 'generation' where students generate their own ideas while solving problems. Student-generated strategies and ideas are sometimes considered an alternative to teacher-, expert-generated worked examples. As discussed above, popular accounts hold that higher levels of support are more suitable for novices and this need gradually reduces with increasing expertise. Thus, there appears to be an intermediate space between novice and expert learning where (along with fading and the introduction of incomplete or incorrect examples) a degree of student generation may be advantageous.

As support is reduced, this often requires students to generate responses or partial responses. Is this effective and how can we identify when this is appropriate?

There were several studies that focused on this 'middle ground' between wholly-supported problem solving and unsupported (or 'discovery') learning, which we summarise below (also see our Working with Schemas section for connected ideas around the elaboration and generation of schemas).

- Glogger-Frey et al. (2015) examined the effects of studying with a worked solution—as opposed to an open problem (inventing)—on student teachers' preparation for a learning-strategy evaluation and for eighth-graders learning about ratios in physics. They found that 'the worked solution prepared learners for learning and transferring their skills to new problems in a better way than the inventing task' (p. 82) and partially explain this with the lower extraneous cognitive load in the worked solution.
- Glogger-Frey et al.(2017) similarly examined the difference between invention and guidance. In this particular study, eighth-grade students either invented twice or worked through worked solutions of the two tasks before learning about ratios in physics from a lecture. They found that, 'Guidance led to less extraneous load. However, self-regulation led to higher transfer because the students devoted more attention to the deep structure of the preparation tasks.' Based on that, they conclude that 'some practice self-regulated outperforms guided preparation for learning from direct instruction' (p.26).
- Likourezos and Kalyuga (2016) compared 'partially-guided or unguided attempts at generating problem solutions as opposed to comprehensive guidance, in the form of a worked example' (p.1). They did not find any differences between the three groups in transfer post-test outcomes. Still, they did find that having fully guided worked examples prior to explicit instruction reduced cognitive load compared to the other conditions without such guidance.
- Chen (2015) discussed the effects of 'worked examples' versus 'generation' in geometry instruction, considering different levels of complexity and element interactivity and effects on students with different levels of prior knowledge in geometry. They found that for materials high in element interactivity there was a worked example effect whereas for materials low in element interactivity there was a generation effect. However, as the level of student expertise rose, so did the generation effect. Thus, they suggest that worked example versus generation effect is dependent on the degree of element interactivity (also, see Chen, 2016).

Timing and sequencing

Does splitting up information reduce cognitive load by allowing piecemeal presentation of material or increase it due to split-attention effects?

An appreciable number of studies in our evidence-base (all rated as medium priority) manipulated the timing or sequencing of the support provided. Rather than being incidental, sequencing forms part of the overall thinking and theory around cognitive load management in complex and high-information tasks. Kirschner and Hendrick (2020), for example, in their presentation of the science for practitioners, discuss the strategy of 'emphasis manipulation' where sub-tasks or sub-concepts are taught individually before being brought together. Similarly, Lovell (2020), again in a practitioner-focused interpretation of the cognitive load theory, discusses 'segmentation' as well as 'sequencing and combination' providing an explanation and practical examples. Some accounts from our practice review discussed 'pre-learning' material that forms part of a larger problem or information space to manage cognitive load when this information is ultimately brought together. An interesting question

with these popular accounts is whether there is a relationship between the principle of split attention (as above), where integrating information in one place leading to lower cognitive load, and sequencing (as here) where splitting up a problem helps manage the overall cognitive load.

For studies in our database relating to sequencing, there were some suggestions that effective sequencing could be effective although we have too little evidence to draw confident conclusions. Below we briefly summarise these studies:

- Jimenez and Saylor (2017) discussed the impact of providing instruction in picture book vocabulary while reading—compared to after reading—the book, building on the idea that if instruction occurs while the book is read it may increase cognitive load. Their experiment involved three- to five-year-old children reading a story that included instruction of six new words either during or after the story. The first approach placed more demand on the children's cognitive resources and therefore the authors argue that it may not be equally appropriate for all children.
- Kester et al. (2005) also discuss the effect of sequential, step-by-step information. Their study discusses the effect of presenting declarative information and procedural information sequentially versus simultaneously both before and during troubleshooting practice of electrical malfunctioning circuits. The stepwise approach was found to free up working memory and facilitate learning.
- Uz-Zaman and Alam (2011) compared learning with pre-lesson assignments using a step-wise approach versus traditional teaching format for maths students. They found that the pre-lesson approach improved students' understanding of maths and indicate that 'reducing working memory demand through pre-lesson assignments leads to understanding' (p. 12).
- Finally, van Zundert et al. (2012) studied the effect of instructing in peer assessment and domain-specific tasks simultaneously compared to providing instruction in the domain-specific task followed by instruction in peer assessment. They found that while 'the final performance of the task (i.e., speed and accuracy in domain-specific skills and peer assessment skills) showed no significant differences […] performance improved more from Phase 1 to Phase 2 in the stepwise condition than in the combined condition'. Based on these results, they argue that 'it might be beneficial to teach domain-specific skills before peer assessment skills' (abstract) in the case of complex study tasks.

One final sequencing-related concept we encountered in the literature was that of 'productive failure'. A small number of studies examined the hypothesis that starting a learning sequence with a problem, even if this leads to failure, might support future learning by helping learners become familiar with the overall problem space (and perhaps also priming them to learn through revealing a gap in their understanding). A recent study in this area by Ashman et al. (2020) examined productive failure for high element interactivity problems with Year 5 primary school students learning about light energy efficiency. Cognitive load theory would predict that in situations involving a 'fairly large number of interacting elements, problem-solving first would overload working memory' (p. 233). They found no support for the problem-solving-first strategy as an effective instructional approach. However, they acknowledge that their experiments involved high element interactivity and that they could therefore not rule out that the problem-solving-first strategy might be useful for tasks with low element interactivity.

Format of information

In what ways does the format and presentation of information affect cognitive load? What are the main variables for teachers to consider?

In a later section, we examine the cognitive theory of multimedia learning and the 'dual coding' of information. However, at this point it is valuable to note the link between cognitive load and presenting and integrating multiple forms of information. This was also a frequent connection made by teachers in our interviews and surveys. Many teachers went from discussing reducing extraneous load in multimedia presentations and other curriculum resources to say that they are trying to present information in line with dual coding theory. There were numerous studies that reflected this connection, looking at how information is presented and the modes in which it is presented in connection with cognitive load (for example, Mousavi, Low and Sweller, 1995; Owens and Sweller, 2008; Haslam and Hamilton, 2010; Richter and Scheiter, 2018).

Many studies examined the impact of information presentation on cognitive load and, consequently, learning. However, these were quite disparate by nature. Below we outline some of the emerging factors relating to cognitive load in our evidence-base.

- **Verbal compared to written instructions** (Liu and Chuang, 2011).
- **Video, interactive, and static presentation** of information (Wang et al., 2020).
- Rekik et al. (2019) and Jarraya et al. (2019) also discussed videos, but in relation to learning of basketball and considering **the complexity and speed of the presentation** of material.
- Reisslein et al. (2015) looked at the effect of changing the colours of mathematical symbols when teaching **novice students** about electric circuits. In a study of high school students, they found that the group who learned with **a combination of colours** achieved higher post-test scores, gave higher liking for the instruction, and had lower ratings of cognitive load than those that learnt using the black font.
- Gnambs et al. (2015) also looked at **the effect of using the colour red in stimulus material**. One hundred and ninety-nine secondary school children were instructed to memorize a short text and subsequently given a knowledge test and a measure of cognitive load. The colour red was manipulated throughout the material. They found that boys were more strongly affected in their test performance by repeated colour exposure than single colour manipulation. For girls, it was the opposite. Similar effects were found for cognitive load.

Several studies, specifically in maths, looked at how **presenting information through equations, word descriptions and pictorial forms** influences cognitive load. Similar themes to the above, around element complexity, interactivity, and student ability, arose from these results, for example:

- Leung, Low and Sweller (1997) explored the cognitive load of **maths equations as opposed to word descriptions**. They hypothesised that equations would impose a heavier extraneous cognitive load on learners because they require a mental integration of notation and meaning. Four experiments were conducted; the results demonstrated that the efficacy of equations and words, and their relative increase or reduction of cognitive load, varied. Efficacy was dependent on the complexity and conciseness of the equation and the extensiveness of verbal information in the word description.
- In a series of studies, Ngu and colleagues (Ngu et al., 2014; 2016; 2018; 2019) compared different approaches to solving percentage change problems from a cognitive load perspective and including a cross-cultural comparison. Similar concepts to those discussed above arose including: **element interactivity, the need to search and integrate information, sequencing and prior knowledge.**

While the evidence we have here is not sufficient for us to draw conclusions about how to present information to optimise cognitive load, several plausible principles are apparent. It seems reasonable

to suggest that considering the impact of different presentational formats, learner cognitive load is valuable and likely to be consequential for learning.

Implementation

What specific activities and strategies can be used to manage cognitive load in the classroom? How do these relate to the concepts identified above?

The above discussion on optimising cognitive load for different students and in different circumstances is mostly at the conceptual level. We close this discussion by reporting some of the detailed strategies teachers are using to manage cognitive load in their classrooms, using direct quotes obtained from our interviews and surveys. Below is a sample of these, grouped under five subheadings:

Sequencing and chunking

- Break coursework or essays down into manageable sections; teach each section explicitly, with worked, narrated examples, exemplars and then scaffolds, which are slowly removed then all elements of piece of work put together for independent practice.
- Scaffolding exam answers: Break the answer into chunks or paragraphs. Complete each section as a class to create a model answer. Students then do the same thing, for the same question but for a different part of the text.
- I am also doing a lot of chunking of content, especially in criminology, so each lesson is clearly signposted and flagged as its own 'part of the puzzle' and linked into the spec to reduce the cognitive load of each unit as there is so much content covered.
- Chunking the subject knowledge helps all the children learn and acquire new skills and knowledge but SEN children really benefit from this.
- Chunking and scaffolding are just basic good primary practice. I would expect all my teachers to demonstrate this.
- Chunking text or vocabulary, creating frameworks and scaffolds to support the understanding of key concepts or language, using models to aid memorisation and build confidence.
- At KS5, I am splitting every lesson into prior learning, new information, and practise to manage cognitive load. Particularly the prior knowledge bit helps to build understanding on foundations that already exist.
- We often use strategies to manage cognitive load, such as modelling writing in chunks—the teacher models a part then the pupils write then the teacher models the next part.
- Stepwise worked examples.
- Cognitive load is considered in the delivery of all lessons, where material is broken down into chunks. Each time new content is taught, it usually requires pupils to build on what they have already discussed in lesson.

Reducing extraneous load

- Thinking clearly about my economy of language during explanations and instructions.
- Reducing cognitive load has supported me in presenting information in a better way and removing unnecessary elements of my PowerPoints used for engagement sakes when actually it may just take up some of their working memory.
- Reduced content on PowerPoint slides. This helps students to focus on key concepts and build knowledge up.
- Decluttered PowerPoint slides.

- Removing unnecessary asides or distractions.
- Not too much writing on PowerPoints. Use of working wall in lessons, partner talk etc.
- We have de-sensitised class bases—not too busy and not full of information that the children do not need.

Scaffolding, modelling, and worked examples

- Scaffolding and modelling of geographical key concepts in combination with dual coding and using graphic organisers, particularly for complex ideas at A-level geography.
- Worked examples support pupils who need support and allow independence but needs to be taught alongside metacognition so pupils understand how to use the examples and how to work towards not needing them.
- Modelling as a way of managing cognitive load … often explore a paragraph as a class before getting them to produce the second paragraph on the same topic.
- Planning tasks with cognitive load in mind is really effective for chunking and scaffolding work—particularly for classes with high SEN need and low reading ages. I am able to deliver challenging content (keeping expectations high) but in small sections and the pupils build the pieces up to like a jigsaw puzzle. This ensures they don't get overloaded, and if they do have difficulties, it is very easy for me to spot the exact stage that is tripping them up allowing me to quickly respond to them.
- Reducing cognitive load by supporting students with selection of relevant information, clear instructions, and modelling of processes to take notes of answer questions.
- Key words for the topic on the wall to ease cognitive load.
- Managing cognitive load through modelling using a visualiser.

Integrated instructions

- Including equations on the question sheets to reduce cognitive load.
- Integrated instructions for science practicals to reduce extraneous cognitive load.

Mixed approaches

- I do, we do, you do—live modelling, eradicating extraneous material from lessons, chunking.
- Strategies to manage overload are useful in the preparation for assessments ... manage cognitive load: 'chunking' subject content—using worked examples, exemplars, or 'scaffolds' are hugely important but often not effectively taught at teacher's training.
- Strategies to manage cognitive load—modelling of concepts with physical props (pipe cleaners, plasticine), stop motion video creation, applying analogies or mnemonics for key points (for example, LORD for left oxygenated and right deoxygenated blood flow around the heart), practical experiments regularly carried out, practical instructions in bullet points, providing key word lists, broken down learning objectives.
- We use mind maps to allow the children to write and draw what they know about a given topic. We use chunking in maths to allow pupils to break down a question into smaller steps. We use scaffolds for children with SEN to allow them access to work and create independence.
- Strategies to manage cognitive overload are necessary in many science topics. I almost always use scaffolds, exemplars, and mnemonics. I always put key words and definitions for the lesson around the room for pupils to refer to if needed and I like to break things down into simple step by step instructions.

What are the challenges of implementing cognitive load management strategies?

Finally, several teachers discussed the challenges they faced in implementing strategies to manage cognitive load. For example, one interviewee participant discussed how some teachers struggle with erroneous examples or models because they worry about introducing misconceptions and because of a perception that teachers 'need to be perfect'. They went on to explain, however, their view that the use of erroneous examples encourages a climate where mistakes are 'okay' and something to learn from.

Other comments relating to the challenges of managing cognitive load from the questionnaire are as follows:

- 'Strategies to avoid cognitive overload are trickier to embed ... I am interested in the cognitive load theory but I often wonder if I am doing it right. Am I achieving the best results? Managing cognitive load is difficult—there's so much to think about!'
- 'Whilst [cognitive load theory] provides some interesting areas for learning it can be difficult to ensure that this is properly differentiated and managed [so as not to] dissuade certain students from learning.'
- 'Cognitive load using pre-prepared core textbook materials from exam boards! It's like they were designed to split attention and overload the visual channel of the working memory.'

Final thoughts on this strategy area

In our systematic review of classroom trials, we concluded that, overall, the evidence is promising and indicative of the value and importance of teachers seeking to optimise cognitive load. Despite the high concentration of studies of worked examples in secondary maths and science, we concluded that considering worked examples and other forms of scaffolding (for example, support and guidance for complex learning or problem-solving spaces) together suggests wider subject and age applicability of the principle, and provides greater confidence in the overall result.

The framing of our conclusion was important. We describe this as follows: 'Our confidence is in the value of optimising cognitive load *rather than a specific strategy for doing so*, or for specific learner needs.' The main results also suggested great complexity and practical challenges for optimising cognitive load in practice. Ecological validity was low for many studies, limiting our ability to generalise the findings to real educational settings.

Perhaps more than any other area we have reviewed, there is a large disconnect between the applied evidence and the prevailing account of managing cognitive load. The applied evidence for managing cognitive load is limited; to the degree it is detailed, it is quite narrow (for example, for use of worked examples in secondary maths and science); mostly it was far more general, allowing conclusions at the general level of 'cognitive load matters' rather than evidential support for specific strategies. In contrast, practical perspectives on cognitive load provide detailed and highly elaborated accounts that cover most subjects, pupil ages, and specific strategies in detail. Our discussion section has stressed that our aim has been to explore the conceptual landscape addressed by teacher perspectives, practice-focused popular accounts of cognitive science, and our wider evidence. Much of this is highly contested and arguably we have treated it too uncritically given the lack of a robust applied evidence. Nonetheless, we hold that there is value in rehearsing and exploring this overall account. The next steps for researchers and practitioners are to now critically and systematically investigate the questions posed and perspectives aired.

B5. Working with schemas

Overview of area

Definitions

The theory that knowledge is organised in the mind as schemas (sometimes 'schemata')—hypothetical mental structures for representing and organising information—is fundamental in cognitive psychology (Yilmaz, 2011). As an individual encounters new learning experiences, their existing schemas are revised and restructured to accommodate the new information learnt. In this section we review all studies in our database that focus on representing and developing schemas. There are a group of learning theories and strategies that seek to activate or represent schemas as a way of presenting connected ideas, identifying a learner's pre-existing knowledge, and then build on this existing knowledge base, often via provision of scaffolds to manage cognitive load and emphasise pertinent information. Working with schemas often involves developing ideas through processes of organisation, comparison, or elaboration.

A related idea is that schemas are, to a degree, personal; therefore, developing a learner's ideas is an active process involving 'construction' of knowledge starting from, and then developing, the learner's pre-existing knowledge and understanding. When students generate their own responses and bring their own conceptions to the fore, this is often described as 'generative learning'. Generative learning is the idea that 'pupils create understanding of what is to be learnt through a process of selecting information, organising it and then integrating it into what they already know.' (Enser and Enser, 2020, p.11). Our sources also discuss the 'generation effect', which is the idea that students are more likely to remember information when they have played an active role in producing it. This idea of active engagement, of working with ideas and gradually developing them over time, was a common aspect of the sources we consulted in this area, for example:

> *'Your grasp of unfamiliar material often starts out feeling clumsy and approximate. But once you engage the mind in trying to make sense of something new, the mind begins to 'knit' at the problem on its own. You don't engage the mind by reading a text over and over again or by passively watching PowerPoint slides. You engage it by making the effort to explain the material yourself, in your own words—connecting the facts, making it vivid, relating it to what you already know.'*
>
> *Brown, Roediger and McDaniel (2014. p.221–222)*

We found many variations on this general account of generative learning in our scoping literature whereby students organise, compare, and elaborate on their ideas. Sometimes this was linked with a view that struggling with a problem, even if unsuccessful, might have benefits for subsequent learning (or 'productive failure'); in other accounts it was linked with the idea that comparison, examples, and analogy might strengthen learning, either by providing more texture or an additional layer to a developing schema, or through bringing into view misconceptions or 'cognitive conflicts'. Other accounts focused on teaching and learning strategies that might help interrogate or elaborate on ideas to strengthen, develop and transfer learning (for example, see Dunlosky, et al., 2013).

Adding to the complexity here, this section connects several ideas we have reviewed. When students generate information from memory, they are engaging in retrieval practice. Where information is

being organised into schematic representations and used to support exploration of a complex learning space, we are managing learners' cognitive load. Where visual representations are used to dual-code (or triple-code) information and are actively engaged with, we are drawing on the cognitive theory of multimedia learning and generative learning theory. When concept maps and knowledge organisation are used to bring together different but related knowledge, this is interleaving. This is a complex space that also connects with ideas around classroom feedback, dialogu,e and interaction that fall beyond the purview of this review.

'Working with Schemas' was not an area we conducted targeted searches for; it was, rather, a group of studies that did not quite fit, or cut-across, other sections while still meeting our eligibility criteria (including being informed by cognitive science). We decided that, although linked, schema development and generative learning represent distinct concepts and principles of value for this review.

Our focus in this section is on the following three strategy areas within this area:

- concept mapping, or knowledge mapping, and organisation;
- schema or concept comparison and cognitive conflict; and
- elaboration and self-explanation.

We did not have sufficient evidence to formally and systematically review the third of these; however, 11 studies relating to elaboration or self-explanation (all medium priority) are explored in the discussion and questions section. This section focuses on reviewing the evidence in the first two strategy areas.

Overview of the evidence-base

Table B5.1: Working with schemas—overview of study priority ratings

Priority Level	Overall rating	Ecological validity	Relevance and definition for focus CS practices	Added value to evidence-base
High	4	10	14	12
Medium	34	40	45	57
Low	49	37	28	18

The review study database contained 87 studies in the Working with Schemas category. Of these, 38 were graded as being of sufficient ecological validity, relevance, and value for inclusion within this analysis of the evidence (high and medium). Four studies scored highly across these criteria and were identified as *potentially* providing strong evidence in this area (high).

This area, like many in this review, included many studies with limitations in ecological validity. There were also some limitations in how tightly studies fit (and provided a test of) our definitions of cognitive science principles and strategies in this area, in large part due to some of the complexities discussed above. As a result, only four studies were rated highly across categories (as a best-fit judgement).

Regarding relevance and definitions, one of the difficulties related to the concept of a 'cognitive science informed strategy'. Our searches were focused on cognitive science principles of retrieval, spacing, interleaving, cognitive load, and dual coding. Our searches also included the general cognitive science terms: 'cognitive', 'brain', 'neuro', and 'learning science' and general memory terms such as 'working' and 'short-term' memory (see Appendix 3 for full details of the literature searches). We did not conduct dedicated searches for concept mapping, comparison, or elaboration. An issue with this

is that the practice of concept mapping, for example, could be said to stem back to more general constructivist learning theories rather than contemporary cognitive science specifically (and this itself is a problematic distinction). In this section, our priority criteria have led us to include studies that refer specifically to cognitive science and schema development (and to related concepts such as cognitive load); but this is arguably a subset of practices in this area as many studies employed (for example) concept mapping strategies but *without* providing a rationale for doing so in terms of cognitive science. Our overall aim is to review the evidence for cognitive science informed practices so, strictly, we are testing, for example, concept mapping that is informed by cognitive science rather than concept mapping *per se*. A clear definition of 'cognitive science informed practice' was a challenging operational challenge for the review, which we discuss at greater length in the review limitations section.

In this area, we have identified two strategies with sufficient evidence to examine the effectiveness of the strategy. These are:

- **concept or knowledge mapping and organisation** (15 studies, of which three are graded as high priority and thereby identified for in-depth analysis); and
- **schema or concept comparison and cognitive conflict** (ten studies, of which one is graded as high priority).

Wider evidence in this area looks at how some of the practical and theoretical aspects of, and ideas around, elaboration and self-explanation in these two areas relate to schema development strategy.

Main findings

Strategy 8: Concept/knowledge mapping and organisation

Concise definition

Concept/knowledge mapping and organisation involves learners creating, being provided with, or engaging with a schematic or organised overview of concepts, knowledge, or information in a learning topic (for example, the water cycle) or object (an encyclopaedia entry about the water cycle).

Full definition and description

Concept/knowledge mapping and organisation involves learners creating, being provided with, or engaging with a schematic or organised overview of concepts, knowledge, or information in a learning topic (for example, the water cycle) or object (an encyclopaedia entry about the water cycle). As explained in the introduction, this was a group of studies informed by cognitive science that related to the schematic nature of knowledge and the potential value of providing or producing representations of this. Some focused on the cognitive load aspect, supporting students to identify the 'deep' structure knowledge and connections between concepts. Other studies emphasised the activation of student's prior schemas and the subsequent development of those existing schemas. Other studies focused on the 'active' of 'generative' development of new schemas, using concept/knowledge maps or organisational approaches to support this.

Selected examples

Examples of this strategy from our database include:

- In Ritchie, Sala and McIntosh (2013) learners studied geographical factsheets and student-created mind maps with explanatory notes (bullet points of the key ideas).
- Milenkovic, Segedinac and Hrin (2014) tested a teaching strategy which was designed to connect three levels of chemistry knowledge (macroscopic, sub-microscopic, and symbolic) within a coherent schema. They assessed the impact of this strategy on cognitive load and performance.
- Ponce, Mayer and Lopez (2013) provided learners with a computer-based tool to translate passages into graphical organisers, identifying causes, problems, and effects. The approach introduced structural and spatial elements to students' engagement with the text using an active (that is, generative learning) approach.
- Karpicke et al. (2014) used mind maps in several ways, across a series of experiments, linked with retrieval practice. In one condition, learners created a mind map from a text, in another, learners retrieved knowledge from a partially completed mind map, and in another they were provided with a question map, which created a mind map using questions as prompts (for example, in a question map on clouds, there were five branches with prompts such as 'describe stratus clouds (shape and colour)' and 'fog is made of what type of cloud?'.
- Merchie and Van Keer (2016) compared student-generated and researcher-generated mind maps. The students engaged with mind maps through a series of exercises involving, for example, retrieving information from the map, explaining relationships between concepts, and connecting the mind map information with prior knowledge, in both individual and group work contexts.

Evidence for this approach

There were 15 studies for concept/knowledge mapping and organisation. Of these, three were graded as high relevance and quality. Full details of all medium and high studies are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed for this strategy are characterised as follows:

- **Pupil age and characteristics**. The age range of students ranged from age 8 to 17. Most of these (12 of 15) fell within the 8 to 14 age range.
- **Location**. Eight studies were from the US, two from Chile, one from Taiwan, Germany, Belgium (Flanders), Serbia, and the U.K., representing a fair but U.S.-dominated sample.
- **Learning areas**. In terms of the subject focus of studies, seven were in science, two maths, one geography, and five were related to the text comprehension, with a literacy or general studies focus. The focus of the science studies (all but one) was also the schematic information of scientific information that had been retrieved *from text*. This area mostly (13 of 15), therefore, is of studies where students learn by extracting information from text and present this in a schematic or organised way. The two maths studies were the slight exception in that they used schemas to represent solution strategies to mathematical problems.
- **Outcome measures**. Nine of the 15 studies made use of a researcher-designed test aligned to the targeted learning content. Two studies used a more general standardised test and four combined study-specific tests designed by the researchers with standardised instruments.
- **Design and delivery**. There were four studies in which instruction was designed and delivered by researchers, three largely involved independent study via computer software, and two were

delivered by teachers but heavily scripted by instructional material or workbooks. Six of the 15 studies were largely, or mostly, delivered by teachers, with varying amounts of training and guidance.

High priority studies in this area

There were three studies in this strategy category that were rated as having high strength and validity of evidence. We conducted in-depth analysis of these studies and have completed a full risk of bias assessment, summarised in the appendix.

Merchie and Van Keer (2016). This study examined the effectiveness of two instructional approaches of mind mapping, used as a meta-learning strategy. The study involved 14 fifth-grade, 15 sixth-grade, and six multi-grade teachers and their 644 students from 17 different elementary schools in Flanders, Belgium. Elementary schools that agreed to participate in the study were randomly assigned to either (a) a condition with researcher-provided mind maps (RPMM), (b) a condition with student-generated mind maps (SGMM), or (c) a control condition. To avoid design contamination effects, teachers within the same school were assigned to the same condition. Classes assigned to the control condition received no explicit text-learning strategy instruction and teachers followed their usual teaching repertoire (unaware of the information provided in the experimental conditions). Teachers in the experimental conditions embedded a specific teacher-directed instructional approach of mind mapping once a week over a ten-week interval in their social study and science lessons during regular classroom hours; 1.5 hours of after-school training was provided for teachers in the experimental conditions. The outcome measure was a recall test produced by the researchers that measured the percentage of correctly recalled text information in a five-step scoring procedure based on previous research.

Key findings. In terms of results, students in the SGMM condition had significantly lower scores after the first phase (pre-test to post-test) than the other two groups, which did not significantly differ. For the second phase (post-test to retention) there were no significant changes for the three conditions. Experimental condition students (RPMM and SGMM) made significantly greater progress from pre- to post-test in applying overt deep-level *strategies*, engaging less in rather surface-level paraphrasing activities, however, this declined by the post-test for the SGMM group. Overall, SGMM appears to have imposed a heavy cognitive load on students and reduced short-term results, but not retention. RPMM results were similar to the control approach; while there was evidence of deeper strategy use, this did not translate into extra learning. Our risk of bias analysis identified some potential selection of reported results from lack of pre-planned analysis and rated the piece as having 'some concerns'. The risk of bias was low in all other areas.

Milenkovic et al. (2014). This study looked at a systematic approach to organising chemistry knowledge at three levels: macro, sub-micro, and symbolic. The study involved 189 high school students, age 16 to 17, in eight classes in two schools in Serbia. In the treatment condition, students followed a systematic approach designed to support them in generating explanations at all three levels and integrating that knowledge in a manner that forms one entirety. In the control group, the teacher presented the information in sequence and raised questions but did not provide the connection with the previously presented content at the macroscopic and symbolic levels. Training in all eight classes was performed by two chemistry teachers; each of the teachers taught two experimental and two

control classes. The outcome measures were researcher-designed, multiple-choice tests for which measurement validation analysis was performed, with satisfactory results.[19]

Key findings. The study found that students in the experimental group accomplished significantly higher average performance on the test (70.73%) than the control group students (37.73%). They also show that the strategy contributes to a reduction in cognitive load and thereby increases the teaching efficiency. Our risk of bias analysis identifies some concerns with the randomisation process, with missing data and attrition and potential selection of reported results through lack of pre-planned analysis. Overall, we rated this as having 'some concerns'.

Ponce et al. (2013). This study examined the use of scaffolded practice in translating text passages into graphic organizers. It employed cluster-randomised sampling of schools to recruit 2,468 fourth-, sixth-, and eighth-grade students in 69 classrooms in 12 schools in Chile. During the selection process, a first group of six randomly selected schools were contacted, and then a second group of six, and finally, a third group of six until twelve schools accepted the invitation to be part of the study. The study had two conditions: a computer-based instruction (CBI) or traditional instruction (TI) group. The software implemented spatial learning strategies to support reading comprehension and writing. These strategies included graphic organizers (for example, cause-and-effect diagrams, comparison matrices, and hierarchy networks), paragraph templates, and text editors to present and summarise text and highlight main ideas. The treatment teachers received 24 hours of training and integrated the CBI applications into the language arts curriculum during one school semester. A standardized test was used to measure reading comprehension and writing. Data was analysed through a statistical multilevel model.

Key findings. The findings showed that students in the CBI group improved their reading and writing skills significantly more than students under traditional instruction, with an effect size of d = 0.30. Our risk of bias analysis raised 'high' concerns with deviations from the intended intervention and 'some concerns' with potential selection of reported results. Overall, this study was rated as having a high risk of bias.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section, we focus on the study outcomes, summarised in Table B5.2. Studies identified as high relevance and quality have been marked with an asterisk.

Table B5.2: Concept or knowledge mapping and organisation—summary of evidence

Study	Focus	Population	Finding
			High Priority Studies
Merchie and Van Keer(2016)*	effectiveness of two instructional approaches of mind mapping used as a meta-learning strategy	14 fifth-grade teachers, 15 sixth-grade, and 6 multigrade teachers and their 644 students from 17 different elementary schools, Flanders	***Negative-neutral for student generated mind maps.*** ***Neutral for researcher provided mind maps*** • Students' evolution in recall performance, control condition students attained a significantly higher free recall score at post-test when contrasted with students from the SGMM-condition. No significant gains were found for students in the RPMM-condition, compared to the control condition students. • Large number of outcome variables reported, with mixed (positive and negative) and mostly statistically insignificant results reported.

[19] Cronbach alpha = 0.91, and good range of item difficulty within tests.

Milenkovic *et al*. (2014)*	Instructional Strategy Based on the Interaction of Multiple Levels of Knowledge Representation	N = 189 high school students, age 16-17, 8 classes, 2 schools. Serbia	***Positive for organising knowledge to make connections*** • Students in E group accomplished significantly higher average performance on the test (70.73%) in comparison to the C group students (37.73%). (d = 1.82, 95 % CI = 1.48, 2.15)
Ponce *et al*. (2013)*	scaffolded practice in translating passages into graphic organizers	2,468 students in 12 schools, 69 classrooms. Chile	***Positive*** • The findings showed that students in the computer-based instruction group improved their reading and writing skills significantly more than students under traditional instructions — yielding an effect size *d* = 0.30 (t(63) = 2.40, p<0.05).
Larger Studies (pupil n > 500) (Medium Priority)			
Romance *et al*. (2017)	multi-year effects of the Science IDEAS model on science and reading comprehension achievement	N = 4,471 grade 3-5 students in 259 classes, in 13 schools.	***Positive*** • For both outcome measures, the Science IDEAS model resulted in higher achievement (ES=+1.08 GE for ITBS Science, SE = 0.18, p<.001; ES=+.57 GE for ITBS Reading, SE=0.14, p<.001) (GE=grade equivalent). Learning gains were also evident for grades 6-7 students taking the intervention in grades 3-4.
Wijekumar *et al*. (2012)	Effect of intelligent tutoring guidance on nonfiction reading comprehension	*N* = 2,643 4th grade 24 elementary schools; 131 classes US	***Positive*** • Students in ITSS condition scored higher on standardised tests *(d = 0.32,* 95 % CI = -0.02, 0.67) and on 4 researcher-developed reading comprehension tests (d = 0.47, 0.43, 0.32 & 0.47) than control students. • NB. "Web-based intelligent tutors … [provide] consistent high-quality modelling, practice tasks, built-in assessments, and strong and customized scaffolding and feedback to the learners." (p.8)
Wijekumar *et al*. (2017)	Effect of intelligent tutoring on recall of expository texts	*N* = 4,001 4th & 5th grade 45 elementary schools; 259 classes, US	***Positive*** • ITSS had a positive (but not always statistically significant) effect in improving both Grade 4 and Grade 5 organised memory structures, and improving reading comprehension Results reported as odds ratios: odds of treatment being low vs middle performance = 0.48 to 0.99 [CI range: 0.39 to 1.37]; odds of being high vs. middle = 1.20 to 2.43 [CI range: 0.83 to 3.10]. (See previous for description of ITSS).
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
Chang *et al*. (2002)	The Effect of Concept Mapping to Enhance Text Comprehension and Summarization	N = 126, 5th grade students, 4 classes, 1 elementary school. Taiwan	***Provides support for correction and scaffolding concept maps, but less support for generating them.*** • For comprehension, group performance ranged from: correction (M = 79.2, SD=13.8) > scaffolding (71.4, 13.0) > generation (69.4, 13.1) > control (66.6, 15.6) (d = 0.85, 95 % CI = 0.31,1.38 for correction) • For summarization, the group scores were correction (M = .057, SD=.017), scaffolding (.050, .011), generation (.045, .017) and control (.040, .012). (d = 1.18, 95 % CI = 0.63, 1.74 for correction). • The map-correction group did better on the post-test than the map-generation group and the control group did. The differences in post-test scores among the scaffold-fading, map-generation, and control groups were not significant.
Fuchs *et al*. (2004)	Effects of schema-based transfer instruction on real-life mathematical problem-solving	*N* = 351 3rd grade 7 elementary schools, 24 classes US	***Positive*** • In Transfer Tests 1: SBTI (d = 3.69) and expanded-SBTI (d = .3.72) outperformed control group • Transfer Test 2: SBTI (ES = 1.95) and expanded-SBTI (d = 2.10) outperformed control group • Real-life problem solving: for Transfer Tests 3 (d = 2.71), and 4 (d = 1.91), expanded-SBTI group outperformed the other 2 groups • Overall, for real life problem solving, expanded- SBTI most effective

Guastello *et al*. (2000)	Concept Mapping Effects on Science Content Comprehension of Low-Achieving Inner-City Seventh Graders	N = 124 low achieving 7th grade students. 1 school. US	***Positive*** • Effect size estimates revealed that concept mapping can be expected to improve comprehension scores of low-achieving seventh graders by approximately six standard deviations over a traditional instructional technique. • When students lack background information on a topic to aid comprehension, the active participation in constructing semantic or concept maps may help students form a cognitive schema to assimilate and relate the new topic information.
Jitendra *et al*. (2009)	Effect of schema-based instruction on mathematical problem-solving	*N* = 148 7th grade 1 school, 8 classes US	***Positive*** • SBI classes outperformed students in control classes on problem-solving measure at post-test and delayed post-test • No differences on standardised maths test
Karpicke *et al*. (2014)	Effect of retrieval practice on recall of science texts	Aged 9-11 yrs 1 elementary school, 4 classes, US Expt.1/2/3 *N* = 94/103/89	***Neutral, more evidence that it is how maps are engaged with that matters*** • Expt.1: no difference in scores between groups (authors suggest due to lack of support/guidance) • Expt.2: effect sizes are small, but hint at a general advantage of concept map activities that provided less support (I.e., partially completed) relative to the condition that provided the most support • Expt. 3: advantage of guided retrieval (retrieval using partially completed concept maps) over restudying, (d = 0.42)
Okebukola *et al*. (1992)	Individual and collaborative concept-mapping	N = 147, 11th grade students. US	***Positive, especially when cooperative*** • Each of the three concept mapping groups in the study outperformed the comparison group. Cooperative groups outperformed individual and comparison. Coop preference + coop work (M=63.2, SD=9.0), individ pref + coop work (M=59.8, SD = 9.3), individ pref + individ work (M=50.0, SD = 10.3) and comparison group (M=46.4, SD=10.9) (d = 1.66, 95 % CI = 1.17, 2.16 for cooperative).
Ponce *et al*. (2018)	Computer-supported learning strategies (inc. graphic organisation)	N = 152 6th grade students selected from four schools located in Santiago, Chile	***Positive, with higher effect for more engagement and generation*** • The graphic organizer group (d=0.86), highlighting + graphic organizer group (d=1.35), and highlighting + notetaking group (d=0.75) each produced higher comprehension test scores as compared to the read-only group, whereas the highlighting group (d=0.15) and notetaking group (d=-0.11) did not. Results are consistent with the idea that filling in graphic organizers is a generative learning strategy, whereas highlighting and typing notes into a textbox are not.
Ritchie *et al*. (2013)	Effect of retrieval practice (with or without mind-mapping) on geographical fact learning	Aged 8-12 1 primary school, UK Expt.1/2 *N* = 109/209 4/8 classes	***Negative, but NB. that this study compares concept mapping to retrieval practice rather than a BAU*** • Overall: retrieval practice is more effective than concept mapping, and is not enhanced when concept mapping is added to it • Expt. 1: children in the retrieval practice group recalled significantly more facts than those in the non-retrieval practice group (d = 0.44, 95 % CI = 0.14, 0.72)), but no effect of concept mapping • Expt. 2: main effect of retrieval practice (n2 = .05), no effect of concept mapping, and with results consistent at both 1 and 5 weeks later
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Hilbert and Renkl (2009) (Expt.2)	Computer-based concept-mapping tool: Self-explaining examples	11th-grade students from 1 high school (N = 76, 20 males, 56 females, mean age: 16.9 years, SD = .78). Germany	***Positive, but requiring self-explanation prompts for full benefit*** • Authors conclude that providing examples of successful mapping instead of practice is sufficient for fostering conceptual knowledge (d= 0.97, 95 % CI = 0.40, 1.54). However, to also attain the benefits of mapping with respect to the acquisition of domain knowledge, the processing of the examples provided has to be supported by self-explanation prompts. Cognitive load was higher in the practice group and the example+prompts group than in the pure example group.

* High priority study identified for in-depth analysis (see above).

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in Table B5.3.

Table B5.3: Concept/knowledge mapping and organisation—quality of evidence assessment (based on the GRADE approach)

Strategy	Concept/knowledge mapping and organisation
Number of studies	There are 15 studies in this area of which three were rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	Fourteen of the 15 studies were randomised experiments. Karpicke et al. (2014) was a quasi-experiment.
Risk of bias	Our risk of bias assessments on the high-quality papers identified some concerns with all papers. One had 'high' risk of bias for both fidelity and missing outcome data (attrition); another raised some concerns with the randomisation process, and missing outcome data. One was low risk of bias in all areas other than the need for pre-planned analysis. We judge, therefore, there to be at least one strong study in this area.
Inconsistency	**Result consistency**. Results in this area were quite mixed, with several neutral or negative results.
Indirectness	**Practice Heterogeneity**. Most studies in this area were focused on the organisation and study of text using concept maps. Two studies used schematic approaches to problem solving in maths and one concerned the organisation and connection of instruction into levels to support more holistic student understanding. The variation mostly stemmed from the different approaches to engaging with the organised material. Some studies used retrieval practice or forms of engagement and self-explanation as a comparison condition; some used these as part of the intervention. All studies in this area share common principles about organisation and development of schematic knowledge, but the specific approaches are highly variable. **Population, measure, and outcome heterogeneity**. Most studies in this group were for late primary to early secondary ages (12 of 15 studies focused on age 8 to 14). There were a variety of geographical locations represented, although the majority were from the U.S. Most studies used a researcher-designed test aligned to content; there were some examples of standardised instruments. **Design and delivery**. There were a mixture of studies designed and delivered by researchers, by computers, and studies delivered by teachers with varying degrees of scope for variation.
Imprecision	**Group sizes**. Most studies in this area were of a small to moderate size (eight with 100 < n < 500) and there were several larger studies (five) as well as two smaller ones (n < 100). High priority and large and medium studies of medium priority provided mixed results: - Milenkovic et al. (2014)*: d = 1.82 (95% CI: 1.48, 2.15); - Ponce et al. (2013)*: d = 0.30; and - Wijekumar et al. (2012): d = 0.11–0.49.
Publication bias	There is no clear evidence of publication bias, although many studies did not report effect sizes for this to be apparent.
Other considerations	**Searches**. As we discussed at the start of this section, we have not conducted targeted searches for 'concept/knowledge mapping/organisation' and so cannot be confident that we have a sufficiently large or representative sample of the literature in this area.
Overall confidence	**Low (++)** Our confidence in the effect estimate is limited: the true effect may be substantially different from our estimate.
Confidence reasons	Key reasons for our low confidence for this result are as follows: - the lack of dedicated searches for the strategy (these studies were identified in general searches for cognitive science informed strategies); - very high variation in specific practices used—the strategy being tested here is very general; and - high and unexplained inconsistency in results.

Summary of findings for this strategy

Main finding. Our tentative conclusion is that concept mapping and organising knowledge can be an effective approach. However, student-generated approaches risk excessive cognitive load or inefficiency (with time spent organising rather than active engagement with material) and benefit from retrieval or self-explanation scaffold.

Estimated impact. There is mixed evidence in this area; overall, the evidence is positive (12 of 17 studies) but there were also two negative results and three neutral results. There was not enough consistency in the results to estimate an effect size for this strategy. The high eligibility and larger studies provide effect size estimates ranging from d = 0.11 to 1.82.

Confidence in impact estimate. Our confidence in the evidence and this finding is low due to high rates or unexplained inconsistency, large variation in the practice in this group, and the lack of targeted searches within this review.

Heterogeneity. As we have indicated in the 'Finding' notes (final column), there appear to be several variables at play, notably, the organisation of knowledge, the engagement with organised knowledge, and the extent to which students have generated or organised the representation (for example, a concept map). The neutral and negative results all provide examples of studies where the level of support, engagement, and generation appear to have been pitched incorrectly given the learners' prior knowledge. Chang et al. (2002) found that correcting mind maps, but not generating them, was effective. Merchie et al. (2016) found some negative (post-test) results for student-generated mind maps but the evidence was neutral for researcher-provided maps. Karpicke et al. (2014) found differences according to level of support, with indicative evidence in Experiment 2 that partially completed concept maps were preferable to complete maps and, in Experiment 3, that retrieval practice using partially completed maps was effective. Similarly, Ritchie et al. (2013) found that retrieval practice was more effective than concept mapping and that adding a concept map provided no further benefit.

Strategy 9: Schema or concept comparison and cognitive conflict

Concise definition

Schema or concept comparison and cognitive conflict concerns teaching and learning activities in which leaners compare two or more contrasting or conflicting concepts or examples with a view to discriminating between these or adopting a given conception.

Full definition and description

Schema or concept comparison and cognitive conflict concerns teaching and learning activities in which leaners compare two or more contrasting or conflicting concepts or examples with a view to connecting, discriminating, or adopting conceptions. Of particular note is the use of 'cognitive conflict' as an instructional device for prompting revision and greater metacognitive awareness of concepts, in particular contrasting intuitive from scientific conceptions. This group also includes comparison and connection of alternative concept representations (for example, graphical and written) or problem-solving procedures (linked to Strategy 3).

Selected examples

Examples of this strategy from our database include:

- In Ziegler and Stern (2016), students were introduced to addition and multiplication problems in parallel, with one on a left-hand blackboard, the other on the right. After teacher instruction, students copied the examples to their books and then worked on worksheets with similar problems. This was compared to a sequential group where students practiced addition for two days and then multiplication for two days.
- Chiu and Churchill (2016) provided students with a software tool that presented graphical, algebraic, and descriptive information showing the representation of algebraic formula on two-way axes. Students had slide bars to alter the coefficient values and thus compare both forms.
- In Day (2015), students were given real-world scenarios with varying levels of concreteness and a simulation (of a polar ice cap). Students interacted with the information, including using sliders to control the simulation. They were asked to compare the concrete cases with the general, abstract taught concepts and identify whether the scenarios were examples. Although on the boundary of our definition, in this study there was a sufficient element of comparing representations and examples (concrete and scientific) as a way of developing schematic understanding.
- Star et al. (2015) examined the use of worked example pairs and encouraged comparison through asking questions:
    - *Which is better?*—for the same problems solved with two different but correct methods;
    - *Which is correct?*—for incorrect and correct methods of solving the same problem;
    - *How do they differ?*—with two different problems solves in two different ways, with a view to increasing understanding of the underlying mathematic concept; and
    - *Why does it work?*—with two correct maths for the same problem, but with the goal of 'illuminating the conceptual rationale in one method that is less apparent in the other method' (p.9).
- Poehnl and Bogner (2013) provoked 'confusion' through identifying and then challenging students' prior 'alternative conceptions' (sometimes called 'p-prims') of scientific phenomena to replace them with the 'scientific conceptions'. They discussed the links between the strategy and cognitive load theory.

Evidence for this approach

There were ten studies for schema or concept comparison (including cognitive conflict). Of these, one was graded as high relevance and quality. Full details of all medium and high studies are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed for this strategy are characterised as follows:

- **Pupil age and characteristics**. The age range of students in this section was fifth grade (age 10 to 11) to age 18, with only one example of a student from the fifth grade. Therefore, students were mostly from the secondary age range, and particularly around sixth to ninth grades (11 to 14).
- **Location**. Studies came from a mix of countries. The most numerous was the U.S. with four followed by Switzerland with two. In addition, there was a single study for each of England, Hong Kong, Nigeria, and Germany.

- **Learning areas**. All studies were of either maths or science. The mathematical examples (six) were largely concerned with contrasted algebraic solution methods (one was about computation estimation). This subgroup, therefore, links strongly to those considered in the interleaving section. The other studies (four) were in science and were concerned with using comparison and cognitive conflict to provide concrete examples (one), develop cognitive conflict and alternative conceptions (two), or both (one).
- **Outcome measures**. Eight out of the ten studies in this area used researcher-designed outcome tests aligned to the content. Two made use of a combination of standardised instruments alongside study-specific tests.
- **Design and delivery**. This area made heavy use of workbooks or computer programmes to deliver the intervention with none or little instructional input. Two studies had instruction delivered by researchers or a mixture of researchers and teachers. Three studies trained regular teachers to implement the intervention.

High priority studies in this area

There is one study in this strategy category that is rated as having high strength and validity of evidence. We conducted in-depth analysis of this study and have completed a full risk of bias assessment, summarised in the appendix.

Star et al. (2015). This study examined the effect of a supplemental comparison curriculum on students' algebra learning. This was an RCT, assigned at teacher level, involving 76 teachers of 1,367 eighth- and ninth-grade students in 56 schools in the U.S. Initially, 141 teachers were recruited but there was high attrition. In the experiment, the 141 Algebra I teachers were randomly assigned to either implement the comparison curriculum as a supplement to their regular curriculum or to be a 'business as usual' control. The desired implementation model involved the use of questions from all three different types of reflection prompts ('understand', 'compare', 'make connections') covered in this order. Teachers would also allow students to engage in a whole-class discussion around the 'make connections' prompts. Teachers were also expected to display and read to the class the take-away page with learning objective. All treatment teachers attended a one-week (35 hours) summer professional development institute, designed and delivered by the research team, to become familiar with the supplemental curriculum materials and the desired implementation model. Teachers had considerable flexibility in selecting which supplemental curriculum materials to use and integrating the supplemental materials with their regular curriculum. Fidelity was low, with many teachers not using the supplementary materials. The results were assessed using two outcome measures: first, a standardized algebra readiness test, the Acuity™ Algebra Diagnostic Readiness Exam (CTB/McGraw Hill, 2007) and second, a researcher-designed multiple-choice test.

Key findings. The results showed no significant relationships between the offer of the intervention and students' overall, procedural, conceptual, and flexibility knowledge. The offer of the intervention was associated with a 0.97% increase in overall knowledge scores, a 2.54% increase in procedural knowledge scores, a 0.88% increase in conceptual knowledge scores, and a 0.63% decrease in flexibility knowledge scores, on average. The large standard errors suggested that the study was underpowered to detect the small effects that were present. Our risk of bias analysis picked on the issues of fidelity and attrition, rating these as resulting in 'high' risk of bias. There were some concerns about the lack of pre-planning of reported results. Overall, this study was rated as having a high risk of bias.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section we focus on the study outcomes, summarised in Table B5.4. Studies identified as high relevance and quality have been marked with an asterisk.

Table B5.4: Schema or concept comparison and cognitive conflict—summary of evidence

Study	Focus	Population	Finding
High Priority Studies			
Star *et al*. (2015)*	effect of an Algebra supplemental comparison curriculum on students' mathematical knowledge	N = 76 teachers, of N = 1,367 8th and 9th grade students, in 56 schools (after high attrition from 141 teachers). US	***Neutral*** • There were no significant relationships between the offer of the intervention and students' overall, procedural, conceptual, and flexibility knowledge. The offer of the intervention was associated with a 0.97 (SE=2.78) percentage point increase in overall knowledge scores, a 2.54 (SE=3.05) percentage point increase in procedural knowledge scores, a 0.88 (SE=2.76) percentage point increase in conceptual knowledge scores, and a 0.63 (SE=3.00) percentage point decrease in flexibility knowledge scores, on average. The large standard errors suggested that the study was underpowered to detect small effects that were present. Fidelity was low, with many teachers not using the supplementary materials.
Larger Studies (pupil n > 500) (Medium Priority)			
There were no larger studies at the medium priority level			
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
Adey and Shayer (1993)	Lessons based on concrete activities, cognitive conflict, metacognition, schema development (bridging[20] of thinking strategies) in science.	N = 424 24 classes of pupils with 'average ability' Year 7/8, Age 11-13, England	***Positive (with some mixed results)*** • Some sub-group results negative but mostly positive. 3 of 4 positive in GCSE science for 12+/11+ Boys (ES = 1.03/-0.22) and Girls (ES=0.19/0.67). 3 groups of 4 positive ES in maths, and all in English.
Day *et al*. (2015)	use of concrete, familiar examples in science	N = 144 7th and 8th grade students, 1 school, US	***Negative for concrete examples*** • Planned comparisons showed that those students in the *Low* Context condition improved significantly between • pre-test and post-test while those in the High Context group showed a small numerical decrease in performance • after training. The researchers conclude that generalisation and transfer was undermined by contextual detail (i.e., concrete examples).
Madu *et al*. (2015)	cognitive-conflict-based physics instruction over the traditionally designed physics instruction on students' conceptual change in heat and temperature	N = 249 senior secondary students from 2 schools purposively sampled from 12 secondary schools. Nigeria	***Positive*** • Students' level of understanding was significantly dependent on the treatment in favour of the cognitive conflict instruction.

[20] i.e., strategies to generalise reasoning to promote transfer.

Poehnl *et al*. (2013)	computer and textbook instruction with involving alternative conceptions (ACs)	N = 398 9th grade students (M age = 14.9 years). Germany	***Negative (but mixed)*** • Short-term results: Only the control group differed significantly from the instruction groups; the instruction groups did not differ from each other. • Long term results: the number of scientific conceptions learned was *lower* in the instruction groups including treatment of ACs, although the effects are not totally clear. These results indicate that activating ACs without encouraging further processing had a negative impact on learning scientific conceptions.
Rittle-Johnson *et al*. (2009)	solving equations, comparing different problem types solved with the same solution method, or different solution methods to the same problem	N = 162 7th- and 8th-grade students, 9 classes from 3 schools US	***Positive for the comparison of different solution methods*** • Comparing solution methods was more effective for supporting conceptual knowledge and procedural flexibility than comparing equivalent equations or comparing problem types. Authors conclude that effective comparisons are not limited to or better if examples share the same solution method. Rather, pairs of problem-solving examples can vary in problem features or solution methods, and contrasting solution methods seems particularly useful for supporting mathematics learning.
Stara *et al*. (2009)	comparison in a classroom context for children learning about computational estimation	N = 157 5th and 6th grade students, 2 schools. US	***Neutral. Some positive, depending on prior knowledge.*** • At post-test and retention test, students who compared were more flexible problem solvers on a variety of measures. Comparison also supported greater conceptual knowledge, but only for students who already knew some estimation strategies.
Ziegler and Stern (2014)	learning elementary algebraic transformations through contrasted comparisons	N = 72, 154 6th graders, in 4 schools. Switzerland	***Positive*** • Both experiments revealed that the students who processed the contrasted, mixed program performed better in differentiating superficially similar algebra principles than the students who received the more conventional sequential teaching materials (Hypotheses 1 and 2). There was a significant decline in performance after several weeks for both groups, with a persistence of the group differences at post-test over ten weeks (Hypothesis 3).
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Chiu and Churchill (2016)	Conceptual variation in algebra teaching, seeing and experiencing different algebraic forms and solving methods simultaneously	N = 70 students, age 16-18, 3 classes, 1 school Hong Kong	***Positive*** • The experimental group significantly attained higher improvements in all areas – graphical properties, concept association, evaluation of solutions and written explanation (d = 1.25, 0.94, 0.95 and 0.80, respectively).
Ziegler and Stern (2016)	Algebra learning using contrasted comparisons	N = 98 6th graders, from 5 classes. Switzerland	***Positive*** • Successful replication of Ziegler and Stern (2014). Students in the contrast group performed better than students in the sequential group on the three follow-up measures (Hypothesis 1): transformation knowledge (d = 0. 94), explicit transformation knowledge (0.62), and misconceptions (0.78). The effects were maintained over time (Hypothesis 2).

* High priority study identified for in-depth analysis.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in Table B5.5.

Table B5.5: Schema or concept comparison and cognitive conflict—quality of evidence assessment (based on the GRADE approach)

Strategy	Schema or concept comparison and cognitive conflict
Number of Studies	There are ten studies in this area of which one was rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	Nine studies are randomised experiments, one was a non-equivalent comparison group design.
Risk of bias	Our risk of bias assessments on the high-quality papers identified high risks of bias for the single study identified for assessment. There were issues related to fidelity and attrition; there were also 'some concerns' with potential selection of results. We cannot, therefore, be confident that any studies in this area have low risk of bias.
Inconsistency	**Result consistency**. The results in this section were mixed. Even disregarding the one study of concrete examples, a third of the results were neutral or negative.
Indirectness	**Practice heterogeneity**. As discussed above, studies in this area all used comparison and often specifically conflicting comparisons as a learning strategy. The one exception was a study of concrete examples. Given the concentration of studies on maths and science, we believe the studies to be sufficiently homogenous for grouping. **Population, measure, and outcome heterogeneity**. Most studies in this area were of secondary age students (mostly 11 to 14 years old). The only subjects represented were maths and science. All studies made use of a researcher-designed test aligned to the content (two additionally used standardised instruments). **Design and delivery**. This area made heavy use of workbooks or computer programmes to deliver the intervention with none or little instructional input. Two studies had instruction delivered by researchers, or a mixture of researchers and teachers. Three studies trained regular teachers to implement the intervention.
Imprecision	**Group sizes**. There were two small studies (n < 100), seven small-moderate studies (101 < n < 500) and one larger study in this area. The only study in this group that we judged to offer precision was Star et al. (2015)* who estimated small, statistically insignificant improvements in scores ranging from 0.63% to 2.54% points on a standardised assessment and a researcher-designed multiple-choice test.
Publication bias	There were too few studies in the group to assess this. There were only two small studies, both positive, compared to more mixed results for medium-sized studies and the high priority study.
Other considerations	**Searches**. As we discussed at the start of this section, we have not conducted targeted searches for 'concept/knowledge mapping/organisation' and so cannot be confident that we have a sufficiently large or representative sample of the literature in this area.
Overall confidence	**Very Low (+)** We have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of effect.
Confidence reasons	Key reasons for very low confidence in this evidence include: - most studies in this area focused on secondary age (mostly 11 to 14 years old); the only subjects represented were maths and science; - this area made heavy use of workbooks or computer programmes to deliver the intervention with no or little instructional input; and - there was high inconsistency in the results, with several negative or neutral results.

Summary of findings for this strategy

Main finding. The mixed results, and the limited evidence, suggest that this is difficult territory to navigate, to ensure students draw accurate conclusions from the material presented to them. Studies such as Ziegler (2014, 2016) suggest that strategy comparison in maths is a promising strategy, a result in line with similar studies in the interleaving section; studies such as Adey and Shayer (1993) and Madu et al. (2015) suggest that comparison and cognitive conflict are potentially powerful in science.

Estimated impact. The evidence is not sufficient for us to estimate an effect size for this evidence group as a whole. We estimate that the highest precision estimate in this group was Star et al. (2015)

who estimated small, statistically insignificant improvements in scores ranging from 0.63% to 2.54% points on a standardised assessment and a researcher-designed multiple-choice test.

Confidence in impact estimate. Our level of certainty in this evidence in very low. There are some promising findings in this section, however, this is a small and varied group with inconsistent results.

Heterogeneity. Overall, evidence in this area is mixed: there were six positive, two neutral, and two negative results. One of the negative results was a study that added contextual detail through concrete examples, which, in retrospect, fitted less well with this group, but we have retained for transparency and alignment with the pre-planned analysis. The negative result relating specifically to alternative conceptions, conflict, and comparison suggests that activating the alternative conceptions without encouraging further processing had a negative impact on learning scientific conceptions. The two neutral results apparently stemmed from: (a) issues with fidelity in a large-scale intervention and (b) students not having sufficient prior knowledge of the compared estimation strategies.

Organisation and comparison of information—overall evidence summary and conclusions

Summary of results

In this section, we have reviewed 25 studies focused on the organisation and comparison of information. We identified two strategies for which we potentially had sufficient evidence to assess effectiveness. Our results for these are summarised in Table B5.6.

Table B5.6: Working with schemas—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Concept/ knowledge mapping and organisation	Fifteen, of which three were graded as high priority.[1]	The evidence was mixed. Overall, it was positive (12/17 results) but the negative studies suggest caution is needed.	Most studies in this group were for students of late primary to early secondary age (12 of 15 studies for age 8 to 14). Most studies were focused on the organisation and study of text using concept maps.	Very low (+)
Schema/ concept comparison and cognitive conflict	Ten, of which one was graded as high priority.[1]	The overall results suggest promise for KS3 science and maths, however, the evidence-base is small and provides mixed results.	All studies were for maths or science, with the vast majority of students in the 11–14 age range.	Very low (+)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Conclusions about strategies in this area

Working with schemas

Our headline conclusions in this area are:

- Concept mapping and, more generally, the comparison of strategies and concepts have wide relevance for education for all learners and subjects with areas of complex and connected information.

- The evidence presented above, and its limitations, means that this area suggests both promise and pitfalls, and raises as many questions as it answers. Each area provides numerous discussion points that we consider further in the discussion and questions section.
- For concept mapping and the organisation of knowledge, there appear to be several variables at play, notably, the organisation of knowledge, the engagement with organised knowledge, and the extent to which students generate or organise the representation themselves (for example, a concept map). Our tentative conclusion is that *concept mapping and organising knowledge* can be effective approaches but that student-generated approaches risk excessive cognitive load or inefficiency (with time spent organising rather than active engagement with material) and benefit from retrieval or self-explanation scaffolds.
- Similarly, for *cognitive conflict and comparison*, the neutral and negative results all provide examples of studies where the level of support, engagement, and generation appear to have been pitched incorrectly given the learners' prior knowledge.

Evidence-informed discussion and questions

About this section

In the review area we brought together several related concepts that did not neatly fit into other areas, and for which we did not conduct targeted searches:

- the organisation and comparison of knowledge; and
- self-explanation, elaboration, and student generation.

Given the lack of targeted searches, the wider evidence we have in this area is, therefore, less extensive than in others. We are also less confident than in other areas of the review that this grouping of concepts is practicable or taps into common mechanisms and practices. Nonetheless, from our practice review literature, data collections, and our main database, we conclude that these ideas belong to an account of cognitive science that future research and practice development might usefully investigate. As we are not confident that common principles are at play, we do not use the general discussion structure used in other sections (theoretical principles, practice variation, and implementation). In this section we:

- discuss the question of whether and how these ideas are linked to each other and to our focus cognitive science concepts; and
- present our wider evidence relating to student self-explanation, elaboration, and student generation: this was an area that we had hoped to review as a main strategy but decided that the weight of evidence was not sufficient for us to do so.

Working with schemas and links to the focus cognitive science concepts

What is more productive, for what, and in what circumstances: (a) organising information, (b) comparing information, (c) engaging with information, or (d) generating information? How do these ways of working with information link to our core cognitive science concepts?

Our very general observation is that there are many types of information and multiple ways of engaging with it. Many of the core cognitive science areas we have reviewed are associated with principles for working with information:

- spaced practice—space out engagement with (the learning or practice of) the information;
- interleaving—engage with different (but related) types of information;
- retrieval practice—recall information from long term memory;
- managing cognitive load—do not overload learners with information; provide them with scaffolds and expert schemas to structure and organise information; and
- dual coding—working memory can process more than one type of information and providing information in more than one mode can support learning.

In this way, we might look to connect the concepts and strategies from within this section to the core strategies.

- **Knowledge organisation**. Is the organisation of knowledge through concept maps and knowledge organisers akin to using and developing expert schemas within worked examples and scaffolds?
- **Comparison and cognitive conflict** involve a comparison of concepts, and especially for two closely related concepts that the learner benefits from discriminating between. Does this draw on the same principles and mechanisms discussed for interleaving?
- **Elaboration and self-explanation**. In the section on retrieval practice we briefly discussed the idea that transfer of learning might be promoted by varying the retrieval approach and content. Moreover, some studies suggest the benefit of 'elaborative retrieval' (Carpenter, 2009). Does elaborative retrieval modify and develop schemas whilst also consolidating them?
- **Generative learning**. One of the principles of the Cognitive Theory of Multimedia Learning (see next section) is that 'learning is an active process of working with information'. Is generative learning a cross-cutting idea that links to other strategies, emphasising the connection of new knowledge with old and the active nature of this process?

In our main results we concluded that there were some encouraging results relating to concept mapping and organising knowledge and cognitive conflict and comparison, but that there were also numerous negative results and complexity in the area. We go on to say that there appear to be several variables at play—the organisation of knowledge, the engagement with organised knowledge, and the extent to which students have generated or organised the representation (such as a concept map). Our tentative conclusion was that concept mapping and organising knowledge can be effective approaches but that student-generated approaches risk excessive cognitive load or inefficiency and benefit from retrieval or self-explanation scaffolds. Similarly, for cognitive conflict and comparison, the neutral and negative results all provided examples of studies where the level of support, engagement, and generation appears to have been pitched incorrectly given the learners' prior knowledge.

Teacher perspectives from the practice review also linked ideas of information organisation, comparison, engagement, and generation. Two excepts from the discussion with Interviewee 3 provide an interesting example:

> *'I was really thinking about the schemas that they could build initially, make everything link back, so I was doing some quite experimental things around reducing plays and novellas to sort of keywords, key quotes, introducing those first, getting their associations and then building on those initial ideas so they could keep linking it back. I stopped reading whole texts and annotating because I find there's nothing for them to remember from that. I stopped doing that and we started to do a lot of pre-reading activities so that there was something for that information to link to.*
>
> *'When you have a whole text to teach, that's when I've really found that the work around schemas was really effective. The other thing I do, I do look at a lot of cognitive theory, but I also think you have to bring in your own subject theory and match the things. So, I was looking around literary theory and the idea of matrices and this idea that the sociological ideas in a text are born from just a very limited number of words in the text. So, for instance in Macbeth there are five words that are repeated so much and nearly every key idea links to the five words. So, I've reduced Macbeth to the five words and I've reduced Christmas Carol to six words and then we started off by building a schema around each word so that when we did look at chunks of the novella they'd already worked out [for example] 'social injustice' and they'd already made that meaningful. They'd come up with their own examples and then they were applying it to the text so that meant there was no learning the text, memorization of the text, it then just became the analysis. So, it really sped up the process. So, every lesson we go back to those six keywords, those initial schemas, and we built—we just keep adding to them.'*
>
> *(Interviewee 3)*

What is connected within this account is the use of ideas relating to (a) schemas,(b) knowledge organisation, (c) generative learning (for example, their own examples), and (d) engagement and active construction (application to, and analysis of, text based on pre-learnt schema). The account also connects these general ideas to subject and topic-specific knowledge. These were ideas we encountered in many areas of the practice review and are explored further below. This has brought us into complex territory, and considerably beyond that tested in the applied evidence. Therefore, here we pose, but do not seek to answer, the following questions:[21]

What is the value of organising knowledge using schematic maps, diagrams, and organisers? Is this more, or less, applicable (and what does it look like) in particular subject areas or for different types of learning content? How important is engagement with organised knowledge? What strategies can be used? What are the practical challenges?

[21] We also refer readers to the short discussion of designing and using knowledge organisers in Sherrington and Caviglioli (2020), p.117.

Elaboration, self-explanation, and student generation

What is the role of elaboration and self-explanation tasks or prompts connected to retrieval practice, organised information, or multimedia information? What strategies can be used? What are the practical challenges?

As noted above, in our main database we had grouped several studies (all medium priority) in the 'working with schemas' area under 'elaboration and self-explanation'. Many combined elaboration and self-explanation as a form of schema development linked with other core strategies, such as worked examples, multimedia information, and retrieval practice. A relevant theory in this area is 'elaboration theory', which links seven strategies, summarised in Kirschner and Hendrick (2020, p.158):

1) an elaborative sequence;
2) learning prerequisite sequences;
3) summary;
4) synthesis;
5) analogies;
6) cognitive strategies; and
7) learner control.

This theory describes how, during learning, schemas develop from simple to complex whereby fundamental ideas gradually become more connected and synthesised with wider ideas. Throughout this process, understanding becomes richer and detailed and learners develop metacognitive control over the topic area.

The studies we have located on elaboration are not empirical tests of general elaboration theory in general; rather, they are focused on the effectiveness of specific, discrete elaboration and self-explanation strategies. Below we provide brief summary of these studies:

- **Berthold and Renkl (2009)** explored the effect of multiple representations (diagrams and equations) on conceptual understanding of worked examples on the topic of probability. They also examined the extent to which learners needed instructional support to utilise these multiple representations. High school students were provided with two types of support: '(a) a relating aid that used colour codes and flashing to help learners see which elements in different representations corresponded to each other on a surface level and (b) self-explanation prompts to ensure that learners integrated corresponding parts in different representations on a structural level' (p.70). Findings showed that the multiple representations themselves did not foster conceptual understanding, but both types of support enhanced it. Self-explanation prompts were found, however, to have conflicting effects on learning outcomes.
- **Wong et al. (2019)** examined middle school students' learning performance with either worked examples (product versus process) or self-explanation (menu-based versus focused). One hundred and twenty-two participants who were acquiring new mathematical skills of calculating the area of a triangle were randomly assigned to one of the four groups and subsequently tested on their performance and cognitive load. In relation to self-explanation, they found that 'menu-based self-explanation prompts may be more beneficial for eliciting accurate self-explanations across prior knowledge levels compared to written focused self-explanation prompts' (p. 21).
- **Hilbert and Renkl (2009)** also looked at worked examples in combination with self-explanation. Their second experiment (with high school students) included a group who studied examples with the additional support of self-explanation prompts. They found that the self-explaining examples

led to better learning outcomes than in the two other groups. However, learning with self-explaining examples also led to a higher cognitive load compared to examples without self-explanation.

- **Fuchs et al. (2015)** tested the effects of teaching fourth-graders at-risk of mathematical learning difficulties to explain their mathematics work. Two hundred and twelve children were randomly assigned to either a control group or one of two variants of a multi-component fraction intervention. The intervention conditions included 36 sessions of 35 minutes of which the last seven minutes consisted of them being taught either to provide high-quality explanations when comparing fraction magnitudes or to solve fraction word problems. They found that the children who received the explaining intervention outperformed those in the word-problem condition. However, the explaining intervention was more effective for students with weaker working memory while the word-problem intervention was more effective for students with stronger reasoning ability.
- Finally, **Willougby et al. (1999)** examined an elaboration strategy with students in Grades 2, 4, and 6. One hundred and thirty-four Canadian students in four schools were assigned randomly to either a verbal elaboration (answer why each fact was true), imagery (create a mental picture), or key-word condition (create a mental picture using provided keywords) and presented with four familiar and four unfamiliar animal story sets. They were then tested for memory of the information. The authors found that whereas the benefits of the imagery strategy were dependent on age (with the older students benefitting from this strategy), the benefits of verbal elaboration were not.

These studies indicate that self-explanation and elaboration may improve learning in some situations but, equally, that they may also risk imposing a higher cognitive load or produce misconceptions. Many teachers in our interviews and questionnaires touched on ideas around self-explanation and elaboration for the purposes of schema development. They discussed the importance of trying to connect the information in meaningful ways ('make connections'), linking new learning to existing schemas ('building on' or 'hooking learning onto' learner prior knowledge), and described the process of schema-building as 'active'. This relates to our final question in this section: the concept of generative learning.

How valuable is it for students to wholly or partly play a role in organising or generating the information? Does this depend on the topic or pupil prior knowledge? What strategies can be used? What are the practical challenges?

As we note above, the idea of generative learning has cut across many areas of this review. Many of the concepts and strategies we have discussed have focused on the content and structure of the information that is (or is not) provided to students during instruction. The idea of generative learning very much focuses on how students can engage with this information.

A practitioner-focused account of generative learning, based on Fiorella and Mayer's ideas in this area (Fiorella and Mayer, 2015; Fiorella and Meyer, 2016) is provided by Enser and Enser (2020). The latter describe generative learning as follows:

> *'What a great deal of educational research adopted by teachers has tended to focus on is the instruction phase of the learning process […] Generative learning considers the learning experience from the point of view not of the teacher, but of the learner. It asks what they should do with the instruction that they have been given to ensure that they are able to truly make sense of it and learn it in a way*

> *that allows them to apply it to new situations in the future […]* Generative Learning in Action *[Enser and Enser's book] is based on a theory of learning that suggests pupils create understanding of what is to be learnt through a process of selecting information, organising it and the integrating it with what they already know.'*
>
> *(Enser and Enser, 2020, p.11)*

As we have noted, this review has been organised in such a way that makes the idea of generative learning one that cross-cuts many other areas and strategies we examine, rather than an area we review in its own right. This is well illustrated by the chapter list within Enser and Enser (2020): (Learning by:) Summarising; Mapping; Drawing; Imagining; Self-Testing; Self-Explaining; Teaching; and Enacting (chapter list from Enser and Enser, 2020, p.3).[22] There is some evidence related to generative learning in cognitive science.[23] We return to the idea of generative learning in relation to the Cognitive Theory of Multimedia Learning in the next section.

[22] We will leave the reader to 'self-generate' the connections to other sections.
[23] See for example, Potts and Shanks (2014).

B6. Cognitive theory of multimedia learning

Overview of area

Definitions

Dual coding theory (a multiple coding theory) observes that information comes in multiple modalities including visual, auditory, and haptic (tactile and motor). Significantly, for dual coding as a theory of learning, there is an 'orthogonal' relation between modes of information in memory (Paivio, 1991)—in other words, our working memory has multiple distinct, yet connected, subsystems. Baddeley and Hitch (1974) proposed a model of the working memory with two distinct components (controlled and monitored by a central executive):

- **a visuospatial sketchpad** that deals with visual and spatial information (for example, the location of a parked car): here we process images 'synchronously', seeing them all at once, their links and location in space; and
- **a phonological loop** that deals with auditory information (for example, the digits of a telephone number): here we process information 'sequentially', experiencing and playing auditory information back to ourselves in a 'loop'.

Several notable implications for teaching and learning stem from this area of research and general description of memory. First, as information can be presented in multiple modes, teachers must consider which mode is most appropriate. This is often driven by the nature of the content to be learnt. However (as studies in this section explore), sometimes teachers are left wondering whether to present information as text, images, diagrams, equations, or some combination of these, and how their choice will affect conceptual understanding and cognitive load. Second, encoding information in more than one mode is thought to strengthen learning. In a seminal experimental test of the dual coding hypothesis, Mayer and Anderson (1991) found that combining spoken words with pictures (an animation) led to more learning than presenting the words before the pictures or presenting either words or pictures alone. By 'dual coding' the information, students could learn and connect information from more than one mode, leading to deeper and more effective learning. Third (an important link to cognitive load theory), because visual and auditory information is processed within connected but separate subsystems of working memory, presenting information in both modes can provide more and richer information *without* overloading working memory. Connections with cognitive load, and whether this holds in applied practical tests, is considered by many studies within this section.

Relevant studies were located through targeted searches for 'dual coding'. We also found that many related studies relevant to dual coding had wider concerns; our broader search terms for cognitive science and memory added to this breadth. We grouped all studies concerned with multiple modes of information into five groups:

1. visual representation and illustration;
2. diagrams;
3. spatial, visualisation, and simulation approaches;
4. audio and images; and
5. animations.

Of these, the first three had a sufficient number of high and medium priority studies for assessment of the evidence. There were around 15 studies each for the fourth and fifth areas, but there were no high priority studies that were highly relevant and performed at sufficient scale in ecological valid conditions. This was disappointing: as per the brief introduction above, these areas—particularly the combination of audio and images—have high relevance to this area. Therefore, while we do not formally review these strategies, we describe these studies in the discussion and questions section with the caveat that—given the lack of strong evidence in this area—this discussion can only be exploratory and its findings indicative.

The other decision to arise from mapping the evidence in this area was our decision to frame this section on the **cognitive theory of multimedia learning** (CTML). As noted, targeted searches were conducted for dual coding, around which this review area was originally planned to be framed. Given the breadth of study strategies and questions, however, we have defined this section according to a slightly broader theory: the cognitive theory of multimedia learning. CTML (Mayer, 2005) builds on ideas of dual coding, cognitive load and generative learning (see previous section) with three core assumptions underpinning the theory: (a) that there are two separate channels for information (that is, dual coding theory), (b) that each channel has a finite capacity, and (c) that learning is an active process of working with this information. CTML also describes organising, selecting, and integrating multimedia information into coherent representations and combining them with prior knowledge (Mayer, 2005). Mayer has, in subsequent publications, set out and described principles of multimedia learning. We return to these ideas in connection with the wider literature in the discussion and questions section.

In summary, for the definitional boundaries of this section, we are concerned with all studies that (a) dual code information or (b) examine (any) strategies for active multimedia learning processing informed by dual coding or CTML theories.

Overview of the evidence-base

Table B6.1: Cognitive theory of multimedia learning—overview of study priority ratings

Priority level	Overall rating	Ecological validity	Relevance and definition for focus CS practices	Added value to evidence-base
High	7	4	30	22
Medium	70	56	79	81
Low	45	62	13	19

The review study database contained 122 studies in the cognitive theory of multimedia learning category. This was the second-largest group in this review, in part due to the wide-reaching nature of visual strategies and our broad definitional focus on multimedia learning along with more tightly focused studies of dual coding. In total, 77 studies were graded as being of sufficient ecological validity, relevance, and value for inclusion within this analysis of the evidence (high and medium). Yet, there were only seven studies that scored highly across these criteria and were identified as *potentially* providing strong evidence in this area (high). We have reviewed these studies in depth and completed a risk of bias analysis.

Overall, relevance to the cognitive science principles was moderate. There were numerous studies of the use of multimedia and visual learning that were quite general or only tenuously linked to the cognitive science in this area. Studies covered a range of areas and a good proportion were rated as adding particular value to the evidence regarding different strategies and their application to different

(subject and age) populations. Surprisingly, given the number of studies in this category, very few studies were rated as having high ecological validity. A key problem here was that the nature of the area lent itself to researchers designing workbooks, instructional computer software, and other specific instructional materials, and these being used to structure and standardise the experiment. There were very few examples of studies delivered or designed with substantial teacher input or materials designed to cover broader curriculum areas (rather than specific topics).

In this area, we have identified three strategies with the potential to provide sufficient evidence to examine the effectiveness of each strategy:

- **visual representation and illustration**—34 studies, of which three are graded as high priority and thereby identified for in-depth analysis;
- **diagrams**—14 studies, of which three are graded as high priority and thereby identified for in-depth analysis; and
- **spatial, visualisation and simulation approaches**—seven studies, of which two are graded as high priority and thereby identified for in-depth analysis.

Wider evidence in this area looks at whether animations or moving images can be more effective than static images, the combination of audio and images, and specific questions about the relationship between cognitive load and multimedia learning.

Main findings

Strategy 10: Visual representation and illustration

Concise definition

Visual representation and illustration involves learners being presented with, or creating, an additional image, picture, or icon that symbolises, illustrates, or represents aspects of the content being learnt.

Full definition and description

Visual representation and illustration involves learners being presented with, or creating, an additional image, picture, or icon that symbolises, illustrates, or represents aspects of the content being learnt. Studies in this group involved the presentation of additional visual information to a task or concept that could potentially be learnt without the visual. In some cases, the visual representation was provided, in other cases produced by the learner. A subset of the this was the use of diagrams, analysed separately (Strategy 11).

Selected examples

Examples of this strategy from our database include:

- pictures representing a (historical) story or labelled historical pictures images (Prangsma, Boxtel and Kanselaar, 2008);
- using simple diagrams (boxes to represent amounts) in maths (Swanson, Lussier and Orosco, 2015);
- the use of circuit diagrams at different levels of abstraction (pictures of lights versus formal symbols for them) in Moreno (2011);

- the use of 2D and 3D images showing chemical structures in Urhahne, Nick and Schanze (2009);
- the use of schematic drawings of cells (with various levels of anthropomorphism) in Schneider et al. (2019);
- including decorative images alongside text—with varying levels of connectedness to the concept being learnt (Schneider et al., 2018);
- using pictures to represent word problems used in primary school maths in Csikos (2012) (for example, pictures of flowers with lines connecting them to bunches of flowers as part of a mathematic problem-solving task); and
- the comparison of text-only, decorative pictures, and representational pictures to accompany fifth grade (age 11 to 12) problem tasks in maths and science in Lindner (2020).

Evidence for this approach

There were 34 studies relating to visual representation and illustration. Of these, three were graded as *potentially* high relevance and quality. Full details of all medium and high priority studies are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed for this strategy are characterised as follows:

- **Pupil age and characteristics**. There was a good age range for studies within this area from upper-primary upwards. KS1 and early years were not represented. Student ages ranged from 7 to 18. There was a good spread across this range, for example, 11 studies included children between the age of 7 and 11.[24] Thirteen studies involved children aged 11 to 15; 14 studies involved children aged 15 to 18.
- **Location**. A range of locations was represented. There were 11 studies from the U.S., ten from Germany, two from the Netherlands, two from Hong Kong, and one from each of Spain, Jordan, Hungary, Finland, Turkey, China, Italy, Australia, and Slovenia.
- **Learning areas**. Fourteen studies in this area (just over a third) are in science, nine were in maths, one in science and maths, three in geography or history, one in music, one in Chinese character learning, three in vocabulary learning and reading, and two concerned learning from a range of texts. This is another area dominated by studies in maths and science, although several others are represented.
- **Outcome measures**. The vast majority of outcome measures were researcher-designed tests produced for the specific targeted learning content (28 of 34). A small number of these reported drawing on previous research or the extant curriculum to create these tests (four) and presumably many more within this group did without explicitly noting this. A small number of studies (four) used standardised tests or items from standardised tests, all doing so alongside or within researcher-designed tests.
- **Design and delivery**. Despite the size of this area, there were scarcely any studies with high ecological validity in terms of their design and delivery. The overwhelming majority of studies in this area can be described as follows: the typical intervention was small, usually conducted in a single school setting; it involved little or no teacher instruction, with only short, scripted instructions, usually focused on provided procedural instructions; it was delivered in a single experimental session, sometimes with prior and subsequent visits to administer tests; the instruction and assessment were delivered via workbooks or computer instructional packages; and interventions were designed and delivered or overseen mostly by experimenters with minimal

[24] Note that there is a small amount of overlap here where studies spanned the age ranges discussed and are double counted.

or no input from teachers. There are only a limited number of exceptions to this characterisation: Csikos et al. (2012), where regular teachers were trained for the programme and given a booklet with lesson plans and materials (high priority study, described in greater detail below); Diana et al. (1997), where the instructional sequence was provided over 12 days and delivered by teachers (with or without supporting the learning with maps); and Kuo et al. (2004) whose study focused on learning Chinese characters spanned over four months.

High priority studies in this area

There were three studies in this strategy category that were identified as *potentially* having high strength and validity of evidence. We conducted in-depth analysis of these studies and have completed a full risk of bias assessment, summarised in the appendix.

Csikos et al. (2012). This study examined the effect of visual representations on mathematical word problem solving. This was a randomised experiment, assigned at class level involving 244 third-grade students in 11 classes in six schools in Hungary. The intervention explored the role of visual representations in word problem solving in mathematics. This included:

- encouraging students to make drawings;
- initiating discussions during problem solving—with 'think aloud' techniques; and
- the use of colourful visual aids.

Teachers received training on the programme and received a booklet containing lesson plans and overhead transparencies with different visual representations attached to the word problems. The programme contained 73 word problems altogether and 20 lessons (four per week) to be taught by children's regular classroom teachers. The control condition was a business-as-usual condition. Students completed their regular mathematics curriculum with no CPD or extra materials. Learning was assessed through an arithmetic skill test comprising 32 items aligned to national core curriculum aims. This was coupled with a researcher-developed word problem test.

Key findings. The study found a small to moderate positive impact of the programme, with an effect size of 0.62 on the word problem test and of 0.20 on the arithmetic test. Our risk of bias analysis identified 'some concerns' with potential section of reported results due to lack of a formal pre-planning of analysis. However, in all other areas the risk of bias was rated as 'low'.

Lindner et al. (2017). This study looked at the effects of representational pictures on maths performance. This was a within-subject experiment involving 401 students with a mean age of 10.7 at three schools in Germany. All students answered 36 manipulated science items that either contained (text-picture) or did not contain (text-only) a representational picture that visualized the text information in the item. Each student worked on both test item types, following a within-subject design. This was a single experimental session facilitated by the experimenter. Students worked individually on computers. The learning was assessed via 36 multiple-choice items adapted from the science framework of TIMSS (see, for example, Mullis, Martin, Ruddock, O'Sullivan, and Preuschoff, 2009). They also measured accuracy, time on task, rapid guessing, and behaviour.

Key findings. The results indicate that (a) RPs improved all students' performance across item positions in a comparable manner (multimedia effect in testing), (b) RPs have the potential to accelerate item processing speed, and (c) the presence of RPs reduced students' guessing behaviour rates to a meaningful extent (that is, a motivational function). Our risk of bias analysis identified 'some concerns' with potential section of reported results due to lack of a formal pre-planning of analysis. In all other areas the risk of bias was rated as 'low'.

Lindner et al. (2020) builds on Lindner et al. (2017) in a study of representational versus decorative pictures on performance in maths and science. Again, this was a within-subject experimental design. The study involved 404 students of mean age 11.4 years at three schools in Germany. All students answered 36 manipulated science items that contained either (a) a representational picture (RP), (b) a decorative picture (DP), or (c) no picture. The study was conducted in two domains: maths and science, producing a three conditions by two domains design. Each student worked on all types of test item types in a single experimental session facilitated by the experimenters. Students worked individually on computers. Again, the learning was assessed through multiple-choice items informed by, or selected from, standardised tests, in this case from the science framework of TIMSS and the German National Educational Panel Study (NEPS; for example, Hahn et al., 2013). Further measures were collected relating to accuracy, time on task, accuracy and item-solving satisfaction, and perceived ease.

Key findings. The results demonstrated that RPs produced greater learning than DPs. RPs enhanced students' performance, perception of ease, and perceived test-taking pleasure in both scientific and mathematics items. RPs increased time on task in mathematics but not in scientific items. DPs had no significant effect on students' performance, test-taking pleasure, or perceived ease. DPs reduced time on task in mathematics items. Our risk of bias analysis identified 'some concerns' with potential section of reported results due to lack of a formal pre-planning of analysis. In all other areas the risk of bias was rated as 'low'.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section we focus on the study outcomes, summarised in Table B6.2. Studies identified as high priority have been marked with an asterisk.

Table B6.2: Visual representation and illustration—summary of evidence

Study	Focus	Population	Finding
		High Priority Studies	
*Csikos *et al.* (2012)	Effect of visual representations on mathematical word problem solving	*N* = 244 3rd grade 6 schools, 11 classes Hungary	***Positive*** • The unbiased effect size for the word problem test was *d* = 0.62 (t = 4.29, p<0.001), and the unbiased d for the arithmetic test is *d* = 0.20 (t = 2.37, p=0.02). • Thus, the intervention program had a small effect on the arithmetic skills, and a notably perceptible (between medium and large) effect size on the word problem test
*Lindner *et al.* (2017)	Effects of representational pictures to testing material on maths performance	*N* = 401 *M* age = 10.7 yrs 3 schools Germany	***Positive*** • The results indicate that (1) RPs improved all students' performance across item positions in a comparable manner (multimedia effect in testing). (2) RPs have the potential to accelerate item processing (cognitive facilitation function). (3) The presence of RPs reduced students' guessing behaviour rates to a meaningful extent (motivational function). Significant positive main effect for adding picture to text (γ1 = 0.30; z = 2.52; p = .012).
*Lindner *et al.* (2020)	Effect of representational (RP) and decorative pictures (DP) on performance in maths and science	Within-subject experiment *N* = 404 *M* age = 11.4 yrs 3 schools Germany	***Positive – for representational but not decorative pictures*** • RPs enhanced students' performance, perception of ease and test-taking pleasure in both scientific and mathematics items. • Furthermore, RPs increased time on task (TOT) in mathematics, but not in scientific items. • DPs had no significant effect on students' performance, test-taking pleasure or perceived ease, while DPs reduced TOT in mathematics items. • Overall: RPs > DPs > Text only, with RP (z = 2.42, p=0.016) and DP (z = 0.28, p=0.780)

Larger Studies (pupil n > 500) (Medium Priority)			
Kutbay *et al*. (2020)	Effect of animation representation on learning electricity	*N* = 855 Aged 11-12 yrs 4 middle schools, 34 classes Turkey	***Mixed results from various comparisons. Spoken better than written.*** • All treatments helped students to develop knowledge of the topic to some extent (pre- to post-test *d* = .34-.81). • No significant difference for abstract animations compared to concrete. • Spoken text + abstract animation outperformed written + abstract animation (*d* = .35, t (171) = 2.31, p = .022) • Spoken text + concrete animation outperformed written + concrete animation (*d* = .35, t (148) = 2.49, p = .014) • While the modality effect held true for middle school students' studying electricity units with a multimedia instruction in real school settings, the signalling and redundancy principles did not.
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
Acha *et al*. (2009)	Effect of verbal and/or visual annotations on vocabulary learning	*N* = 135 Aged 8-9 years 3 primary schools Spain	***Negative*** • Sig main effect of annotation group • 'word-only' groups showed a higher percentage of recalled words than 'word and picture' group and 'picture only' group', both in the immediate and the delayed post-tests • Performance in the 'picture-only' and 'word and picture' groups similar at post-test (d = 0.04, 95 % CI = -0.68, 0.75)
Aldalalah and Fong (2010)	Effect of audio/ image/text on music theory learning among students of different music intelligence levels	*N* = 405 3rd grade 6 primary schools Jordan	*Not test of added image – included for comparison* • Main effect of music intelligence level (high = high) in all 3 treatment groups • The audio-image mode of presentation produced greater learning than text-image and audio-text-image, but especially in pupils with low music intelligence
Ardasheva *et al*. (2018)	Effect of representation and glossary label visuals on science outcomes	*N* = 174 Aged 11-13 years 1 high school US	***Neutral*** • Regardless of English proficiency, English Learners in both treatment and control conditions performed similarly on reading comprehension (d = 0.92) and triggered interest (d = 0.20) measures, with a trend in means favouring the no-visuals, control group
Berends *et al*. (2009)	Effect of illustration type on arithmetic performance	*N* = 135 (divided into poor vs good arithmeticians) *M* age = 9 years 17 primary schools The Netherlands	***Negative*** • Main effect of illustration type on accuracy and time to complete (effects similar between poor and good arithmeticians): • In sum, accuracy of performance dropped when the children were forced to look at the illustrations to find the necessary information (''essential'' illustration). • Speed of performance dropped across types of illustrations, with ''bare'' illustrations being solved fastest, ''essential'' illustrations requiring the longest time for problem-solving, and the ''useless'' and ''helpful' 'illustrations requiring about the same time in between ''bare'' and ''essential'' types.
Chiu *et al*. (2017)	Effect of visual aids and learner expertise on higher order mathematics thinking	*N* = 123 Aged 16-18 years 1 school, 5 classes Hong Kong	***Positive for novice*** ***Negative for advance*** • Results supported the expertise reversal effect for developing understanding (I.e. less structured) but not for remembering (I.e. more structured): • For understanding, novice students who received the aid outperformed those novices who did not receive the aid. In contrast, advanced students who received the aid performed less well than advanced learners who did not receive the aid • All students effectively remembered what they had seen or learned from the multimedia presentations with or without the aid

Diana *et al*. (1997)	Effect of geographic maps on geography fact learning	*N* = 258 6th grade 4 schools, 13 classes US	***Positive*** • Immediate tests of recall indicated that maps facilitated learning related verbal information regardless of prior knowledge of the subject-matter or aptitude level. • Delayed tests of recall suggested that these effects weakened over time
Edens *et al*. (2001)	Effect of pictorial representation on concept learning in science	*N* = 184 4th and 5th grade 1 elementary school US	***Mixed*** • No effect of group on factual post-test scores • Sig effect of group on conceptual understanding: generative drawing produced significantly superior performance in comparison to the writing group (control) (d = 0.66, 95 % CI = 0.29, 1.03). The copying group did not differ significantly from the writing group (control) (d = 0.10, 95 % CI = -0.25, 0.45) • Drawing quality important: students in two picture groups who accurately represented the conceptual knowledge in the text scored significantly higher on the post-test
Gambrell *et al*. (1993)	Effect of mental imagery and illustrations on reading performance	*N* = 120 4th grade 3 elementary schools US	***Positive*** • Free recall: all 3 experimental groups statistically superior to the control group, but mental imagery + illustrations group best of all (d = 1.30, 95 % CI = 0.74, 1.86) • Cued recall: as with free recall, performance of the subjects in the imagery + illustrations group was statistically superior to that of the subjects in the other three treatment conditions (d = 1.34, 95 % CI = 0.78, 1.90)
Gerjets *et al*. (2009)	Effect of representational format and learner control on maths performance in a hypermedia environment	2 experiments 6 high schools Germany Expt.1a: *N* = 118 *M* age = 16.6yrs Expt.1b: *N* = 78 (in addition to 118 above) M age = 16.6yrs	***Neutral / mixed*** • Expt1a: little evidence for the benefit of multimedia design principles for hypermedia when a low level of learner control • Expt.1b: high level of learner control positively affected the effectiveness of instruction only with regard to intuitive knowledge but was at the same time accompanied by large increases in learning time, thereby rendering the instruction inefficient. • Unexpectedly, effects of learner control were not moderated by students' prior knowledge
Homer *et al*. (2010)	Effect of iconic representations on chemistry learning	*N* = 186 Aged 10-13 yrs 1 middle school, 1 high school US	***Positive for low prior knowledge*** ***Negative for high prior knowledge*** • 3-way interaction was found between prior knowledge, age group and icons: ○ Icons were effective for all middle-school students and for high school students with low prior knowledge, but were not effective for high school students with high prior knowledge • Indicate that the expertise reversal effect can be mediated by cognitive development and other factors, not just domain specific prior knowledge
Kiili *et al*. (2006)	Effect of student-generated illustrations on learning about the human immune system	*N* = 167 Aged 10-12 yrs 1 elementary school Finland	***Positive*** • Students performed better on a retention test when they generated their own illustrations by drawing and when explanations were presented as animations, compared to students who received only textual material or generated illustrations from images offered (d = 0.998, 95 % CI = 0.52, 1.48)

Leopold *et al*. (2015)	Effect of instruction representation on science learning	Expt.1: *N* = 112 11th grade 2 high schools Germany Expt.2: *N* = 55 10th grade 1 school Germany	***Positive*** Expt. 1: • The text-picture group performed better than the text-only group on the comprehension test (d = 0.73), the transfer test (d = 0.58), and the referential connections test • The separation instruction was detrimental to the above multimedia effect (d = 0.79 for transfer & 0.47 for comprehension) • Plus, adding explicit integration instructions (I.e. integration group) had no additional effect) Expt.2: • Findings largely replicated the above, but demonstrated applicability to a different learning strategy (summarising instead of identifying)
Leutner *et al*. (2009)	Effect of drawing and imagery on learning science from text	*N* = 112 M age = 16.1 yrs 1 high school Germany	***Mixed – but note that student generated images negative, imagined images positive*** • Results indicate that drawing pictures, mediated through increased cognitive load, decreased text comprehension (d = 0.37), whereas mental imagery, although decreasing cognitive load, increased comprehension, but only when students did not have to draw pictures simultaneously (d= 0.72) • No evidence was found that the effects were moderated by domain-specific prior knowledge, verbal ability, or spatial ability.
Moreno *et al*. (2011): Exper-iments 1 & 2 only	Effect of concrete and abstract visual representations on learning about electric circuits	Expt.1: *N* = 71 *M* age = 13.7 yrs US Expt.2: *N* = 128 M age = 15.4 yrs US	***Positive*** • Generally, Group abstract+concrete (AC) showed an advantage across the experiments: ○ Group AC outperformed Groups A and C on problem-solving practice in Expts 1 and 2 and also outperformed Group C on transfer across both expts • Further, Group A outperformed Group C on transfer in Expt 2
Prangsma *et al*. (2008)	Effect of collaborative construction of representations on learning in history	*N* = 143 Aged 12-13 yrs 3 high schools The Netherlands	***Neutral (no long-term gain after initial positive results)*** • The Timeline group performed significantly better on the post-test than the Text group (d = 0.72, 95 % CI = 0.17, 1.28). On average, the scores of the Visual group on the post-test were 1.75 higher than the Text group scores, but this difference was not significant (d = 0.41, 95 % CI = -0.10, 0.92). But there were not find significant differences between conditions for the retention test scores, which means there is no difference in long-term effects.
Purnell *et al*. (1991)	Effect of technical illustrations on geography comprehension	5 experiments 1 school in Australia Expt.1/2/3/4/5 N = 75/130/24/25/25 Age 14-16/15-19/16-19/ 15-17/14-15	***Positive*** • Experiment 1 – comprehension of a text is not improved by the presence of a technical illustration with content related to but not overlapping the • content of text. • Experiments 2, 4, and 5 – presentation of the same conceptual and spatial content in both an illustration and text results in better comprehension than simple repetition in either text or illustration. • Experiments 2, 3, 4, and 5 – when it was possible to present essentially the same content either in an illustration or as text, comprehension was superior for the illustration.
Richter *et al*. (2018)	Effect of 'signalling' text and/or pictures and prior knowledge on chemistry learning	*N* = 127 Aged 14-17 yrs 3 high schools, 7 classes Germany	***Positive for lower prior knowledge students*** ***Negative for higher prior knowledge students*** • NB. Signalling highlights correspondences between verbal and pictorial information by means of multimedia integration signals (e.g., colour coding, deictic references, or labels) • Results corroborate authors main assumption that prior knowledge moderates the signaling effect in that signals help lower prior knowledge learners but hinder learning in more advanced students, thereby suggesting a full reversal of the signaling effect for high prior knowledge learners.

Schlag *et al*. (2011)	Effect of a strategy to support learning from illustrated texts (about honeybees)	*N* = 133 *M* age = 11.59 yrs 2 schools, 5 classes Germany	*Not test of illustration – but of strategy for engaging with it.* • The students in the Strategy Group performed significantly better with respect to all three types of knowledge: factual knowledge (d = 0.80, 95 % CI = 0.45, 1.15), conceptual knowledge (d = 0.64, 95 % CI = 0.29, 1.00), and transfer knowledge (d = 0.57, 95 % CI = 0.22, 0.92)
Schneider *et al*. (2018): Expt. 3 only	Effect of decorative pictures on learning from texts	*N* = 162 *M* age = 17.6 yrs 1 secondary school Germany	***Positive for connected and positive images*** ***Negative for negative and tenuous*** • Based on results, pictures with a weak connectedness and negative content can be seen as detrimental to learning (d = -0.61, 95 % CI = -1.10, -0.10) and pictures with a strong connectedness and positive content as conducive to learning (d = 0.77, 95 % CI = 0.26, 1.28)
Schneider *et al*. (2019)^	Effect of anthro-pomorphism on learning with media in science	3 Experiments 1 secondary school Germany Expt.1: *N* = 87 *M* age = 11.4 yrs Expt.2: *N* = 148 *M* age = 14.1 yrs Expt.3: *N* = 162 *M* age = 17.6 yrs	***Positive for anthropomorphic despite increasing cognitive load*** • Overall, all three experiments and the • subsequent meta-analysis revealed that an increased degree of anthropomorphism led to an increase in learning, despite increasing extraneous cognitive load. • Anthropomorphism leads to an increase in a perception of aesthetics, intrinsic motivation, and perceived understanding. Authors conclude that results show that the costs of an increased ECL can be offset by rather learning-facilitating factors like motivation and aesthetics.
Starbek *et al*. (2010)	Effect of animation on genetics knowledge and comprehension	*N* = 468 Grades 12-13 (Age 18) 4 high schools Slovenia	***Positive*** • Results for acquired knowledge showed the highest achievement in the text and illustration (d = 1.00) and the multimedia groups (d = 0.94), then in the traditional study group. The text only group showed considerably lower scores. In tasks assessing improved comprehension, the highest achievement was again found in the text and illustration and multimedia groups. Very similar scores to the multimedia group were found in the traditional study group. Again, the text only group exhibited the lowest scores.
Swanson *et al*. (2015)	Effect of verbal and/or visual strategies on mathematics problem solving	*N* = 192 2nd and 3rd grade US	***Positive for students without maths difficulty*** ***Negative for students with maths difficulty*** • In general, treatment outcomes were higher when WM demands were set to a high rather than low level. When set to a relatively high WM demand level, children with maths difficulties performed significantly better under visual-only strategy conditions and children without MD performed better under verbal + visual conditions when compared to control conditions.
Urhahne *et al*. (2009): Study 2 only	Effect of 2D or 3D computer simulations on understanding chemical structures	*N* = 155 *M* age = 16.2 yrs 5 secondary schools, 8 classes Germany	***Positive for conceptual (or 3D vs. 2d)*** ***Neutral for factual knowledge*** • The group with 3D-simulations outperformed the group with 2D-figures in conceptual knowledge. • However, there was no difference in factual knowledge between students who learnt with 3D-simulations and students who learnt with 2D-figures.
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Barbieri *et al*. (2019)^	Intervention using number lines and 'incorporating key principles from the science of learning'.	*N* = 51 2 middle schools 6th Grade US	***Positive*** • The experimental group demonstrated significantly more learning than the control group from pre-test to post-test, with meaningful effect sizes on measures of fraction concepts (*g* = 1.09), number line estimation as measured by percent absolute error (g =.85), and magnitude comparisons (*g* = 82). These improvements held at delayed post-test 7 weeks later

Chiu *et al*. (2016)^	Effect of representation on algebra concept learning	*N* = 78 Aged 16-18 years 1 school, 3 classes Hong Kong	***Positive*** • The experimental group performed significantly better than the control group on algebra learning achievement. The results also showed that only the experimental design with the addition of the instructional approach resulted in higher-order mathematical thinking skills and improved procedural skills of the students (d = 0.67, 95 % CI = 0.19, 1.52)
Cohen *et al*. (2012)	Effect of picture representation on science vocabulary learning	*N* = 89 *M* age = 10 years 2 schools, 5 classes US	***Positive*** • Students in the imagery intervention groups (Picture Presentation, Image Creation—No Picture, and Image Creation—Picture) scored higher on the outcome measures than a word only condition at both immediate and delayed recall.
Kuo *et al*. (2004)	Effect of mnemonic representation on learning Chinese characters	*N* = 92 Age/grade unclear 1 high school, 4 classes US	***Positive*** • Participants who generated their own mnemonics demonstrated higher post-test performance than those in visual mnemonics (d = 0.73), verbal mnemonics (d = 0.75), and translation groups (d = 0.97) (but not the dual coding mnemonics group) • Subjects in the dual coding group scored higher than those in the translation group (d = 0.87, 95 % CI = 0.14, 1.48)

Liu *et al*. (2020)	Effect of dual-coding based computer-assisted learning on vocabulary learning	*N* = 88
		8th grade (~15 yrs)
		1 high school
		China
			Positive
			• Benefit of computer-assisted learning over control:
			  ○ Improved students' learning attitude
			  ○ Enhanced students' vocabulary learning effectiveness

Mason *et al*. (2013)	Effect of concrete and abstract illustrations on science learning	*N* = 59 *M* age = 16.4 yrs 1 high school Italy	***Positive*** • overall, the readers of the text illustrated by either a concrete (d=.55) or an abstract picture (d=.57) outperformed text-only condition. No stat sig differences between the two illustrated texts • Eye-fixation data revealed: abstract illustration promoted more efficient processing of the text, readers made a greater effort to integrate verbal and pictorial information
Moreno *et al*. (1999)	Effect of multimedia-supported metaphors on arithmetic concept learning	2 experiments, US Expt.1: *N* = 60 6th grade 1 elementary school, 2 classes Expt.2: *N* = 26 10th grade 1 elementary school	***Positive for high-achieving students and difficult problems*** ***Negative for low-achieving students and easier problems*** • Expt.1: compared to the single representation group, the multiple representation (visual/verbal/symbolic) group (a) showed a larger pre-test-to-post-test gain for high-achieving students (d = 1.10, 95 % CI = 0.37, 1.84) but not for low-achieving students (d = -0.47), (95 % CI = -1.24, 0.29), (b) showed a greater gain on difficult problems but not easy problems (d = 0.64, 95 % CI = 0.11, 1.17), (c) learned faster during training • Expt.2: high spatial ability students in the MR group outperformed low spatial ability students on pre- test-to-post-test gain (d = 0.93, 95 % CI = 0.36, 1.51)

* High priority study identified for in-depth analysis; ^ = study included for more than one strategy.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in Table B6.3.

Table B6.3: Visual representation and illustration—quality of evidence assessment (based on the GRADE approach)

Strategy	Name
Number of studies	There are 34 studies in this area of which three were rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	All studies are randomised experiments.

Risk of bias	Our risk of bias assessments on the high-quality papers found low risk of bias in all areas for all three studies except for pre-planning of results. We felt that this issue largely stems from an expectation built into the risk of bias tool rather than any particular concern we identified with the studies. We judge, therefore, there to be at least three strong studies in this area.
Inconsistency	**Result consistency**. Result consistency was low. Looking at the results with a crude comparison of the effect of adding or increasing engagement with visual aids, there were 24 positive findings, five neutral or mixed findings, and nine negative findings. As we have touched on in the opening of this section and through information on individual studies, one might distinguish between types of images (decorative and representational) and their suitability for different types of learning and students (for example, high or low prior attainment). As these are factored in, many of the negative results can be said to be consistent with the theory. We report the overall ('crude') results here but return to discuss the moderating factors below and in the discussion and questions section.
Indirectness	**Practice heterogeneity**. With such a large number of studies in this area, there was inevitably variation in the specific strategies and the teaching and learning intentions to which they were applied. All studies tested the impact of adding an image, usually as compared to a text-only condition. Most of the variation stemmed from (a) the role and conceptual content of the image, (b) the format and content of other modes of information and how complementary these were, and (c) how the image was engaged with (including student generation). **Population, measure, and outcome heterogeneity**. As discussed above, the age range of students was 7 to 18, with a good spread within this range. Note that there is no evidence for children younger than this. Studies came from a range of locations and represented a range of subjects; 24 of the 34 studies, however—two-thirds—were of maths and science. **Design and delivery**. Despite the size of this area, there were scarcely any studies with high ecological validity in terms of their design and delivery. As described above, these were mostly one-off, researcher-delivered sessions heavily structured through workbooks or computer software.
Imprecision	**Group sizes**. There were 13 studies involving fewer than 100 students. There were 25 studies of small-moderate size with between 101 and 500 students. There was only one study (N = 855) larger than this. In the small to moderate group, there were four studies with N between 400 and 500, but the majority (19) had fewer than 200 participants. Effect size reporting is, in general, limited and unclear. Estimates provided with confidence intervals were: - Acha et al. (2009): d = 0.04 (95% CI: -0.68, 0.75) - Edens et al. (2001): generative, d = 0.66 (95% CI: 0.29, 1.03); copying, d = 0.10 (95% CI: -0.25, 0.45); - Gambrell et al. (1993): mental imagery + illustrations, d = 1.30 (95% CI: 0.74, 1.86); imagery + illustrations, d = 1.34 (95% CI: 0.78, 1.90); - Kiili et al. (2006): d = 0.998 (95% CI: 0.52, 1.48); - Prangsma et al. (2008): timeline vs. text, d = 0.72 (95% CI: 0.17, 1.28); visual vs. text, d = 0.41 (95% CI: -0.10, 0.92); and - Schneider et al. (2018), Expt. 3 only: weak connectedness and negative content, d = -0.61 (95% CI: -1.10, -0.10); strong connectedness positive content, d = 0.77 (95% CI: 0.26, 1.28).
Publication bias	There is a suggestion of publication bias, with a large proportion of positive results for smaller studies.
Other considerations	This area has a large number of medium-sized studies.
Overall confidence	**Low (++)** Our confidence in the effect estimate is limited: the true effect may be substantially different from our estimate.
Confidence reasons	Key reasons for downgrading confidence to low for this strategy group evidence are: - ecological validity was low: studies were mostly one-off, researcher-delivered sessions heavily structured through workbooks or computer software; - result consistency with low—with a wide range of effect sizes and many negative and neutral results; - studies came from a range of locations, representing a range of subjects, although 24 of the 34 studies were of maths and science; and - there are apparent distinctions needed to (*post hoc*) make sense of these results, between types of images (decorative and representational) and their suitability for different types of learning and students (for example, high or low prior attainment).

Summary of findings for this strategy

Main finding. The evidence suggests that visual aids are most helpful during learning but frequently have no effect, and can sometimes be harmful. The results suggest a need to distinguish types of images (decorative and representational) and their suitability for different types of learning and students (for example, high or low prior attainment). There was a range of subjects, albeit with a disproportionate number of maths and science studies.

Estimated impact. Overall, the evidence points to a small positive effect but with a range spanning from a large positive effect to a small negative effect (see GRADE table for indicative results). We cannot provide an estimate of effect with confidence.

Confidence in impact estimate. Our level of certainty in this finding is low. Key reasons for this include low ecological validity, low result consistency—with several negative effects, and imbalance in the subjects represented.

Heterogeneity. As we note above, a slightly more nuanced interpretation of the theory would hold that the impact of images would depend on their decorative or informational content, their role and centrality within the learning, the format and content of other modes of information and how complementary these were, how the image was engaged with (including student generation), the student prior knowledge, the overall cognitive load, and more. The evidence is not sufficient for us to make these distinctions and reach robust judgements on the effect sizes for subgroups and their impact on different learning outcomes and populations. We return to this question and start to tease apart some of these factors in the discussion and questions section.

Strategy 11: Diagrams

Concise definition

The use of diagrams involves learners being presented with or creating an image that represents or organises learning content or process information schematically.

Full definition and description

The use of diagrams involves learners being presented with or creating an image that represents or organises learning content or process information schematically. In most cases, the diagrammatic representation is an object of learning in its own right. The diagram goes beyond illustration and decoration to represent relevant concepts or phenomena and, additionally, how they are organised or structured.

Selected examples

Examples of this strategy from our database include:

- Chen (2019) investigated learning from scientific ideas encoded in either diagrams or summaries. They examined this in relation to students' level of expertise in the topic and considered perceived cognitive load.
- Chu (2017) provided diagrams that represented quantities and terms within equations. This was highly similar to Swanson, Lussier and Orosco (2015) (example provided in Strategy 10) but with more complex learning material. In Chu (2017), we judged the image to be a diagram

as it depicted structure and connections between representations whereas in Swanson et al. (2015) we judged the image to represent singular concepts. These provide a useful case for which to operationalise the distinction between diagrams and images.

- Cromley et al. (2013b) compared multiple diagram instruction methods: students generating explanations from a textbook diagram and prompts, students completing a diagram with graphic elements, or students completing a diagram with text labels and explanations.
- Diagrams of the water cycle in Coleman, McTigue and Dantzler (2018) with several versions including: labels, labels and process information (for the water cycle), and integration of the textbook information and explanations about the water cycle.

Evidence for this approach

There were 14 studies specifically focused on using diagrams (a subset of the previous strategy). Of these, three were graded as *potentially* high relevance and quality. Full details of all medium and high priority studies are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed for this strategy are characterised as follows:

- **Pupil age and characteristics**. Most studies in this area (12 of 14) were for students between the age of 12 and 16. Two studies were for third- and fourth-grade students (age 8 and 9) and one that, in a sub-experiment, included students with a mean age of 17.
- **Location**. Ten of the 14 studies were conducted in the U.S. The other four were from Australia (two), the Netherlands (one), and Australia (one).
- **Learning areas**. This is another area where studies are heavily concentrated in maths and science learning. There were five studies of maths, including algebra, geometry, and problem solving. There were eight studies of science, including biology, physics, chemistry, and more general science. There was one study of technical illustrations in geography.
- **Outcome Measures**. Thirteen of the 14 studies used a researcher-designed test as a key outcome measure. Several of these were based on the existing school curriculum or previous research. There was one study that made use of a mixture of standardised curricular and psychometric tests.
- **Design and delivery**. Most of the studies (11) were delivered via workbooks or computer programmes, with minimal teacher input. Some of these were administered by teachers, but with minimal instructional input. Others were carried out or facilitated by the experimenter. There were three studies where teachers received professional development and resources (such as workbooks) and then implemented these in their classroom.

High priority studies in this area

There were three studies in the diagrams strategy category identified as *potentially* having high strength and validity of evidence. We conducted in-depth analysis of these studies and have completed a full risk of bias assessment, summarised in the appendix.

Bergey et al. (2015). This study looked at the effects of diagrams versus text for spaced restudy on biology knowledge and comprehension. This was an experimental design with conditions assigned at an individual student level. The study involved 128 students of mean age 14.9 in 15 classes in one school in the U.S. The diagram-based restudy (treatment) condition was a sequence of warm-ups that addressed key concepts presented in diagrams from the students' biology textbooks. Each warm-up directed students to examine a specified diagram in their textbook to answer the questions. In addition, each warm-up included a diagram decoding tip that explained the use and importance of a relevant diagram convention (for example, captions, labels, arrows, and colour coding). The warm-ups

asked students to answer two questions that required the use of a specific diagram from the textbook. The business-as-usual condition was a similar sequence of half-page, text-based warm-ups addressing the same concepts as the DBR condition. Over four weeks (13 instructional periods), teachers distributed individually labelled warm-ups to students as they entered the classroom each period. Warm-ups were designed (by the researchers) to be completed during the first five minutes of each class period, in line with existing instructional practice at the high school. The outcome measures were a curriculum-based test of basic biology knowledge, a biology diagram comprehension measure, a test of basic geology knowledge, a geology diagram comprehension measure, and one measure of spatial ability.

Key findings. The study found equal, significant, and moderate- to large-sized progress in *both* conditions on biology knowledge (d = 0.46–0.51), biology diagram comprehension (near transfer: d = 0.31–0.63), and geology diagram comprehension (far transfer: d = 0.59–0.67). In other words, diagram-based warm-ups were not more effective than text-based restudy. Whether delivered in a text-based format or a diagram-based format, restudy warm-ups were associated with significant growth in biology knowledge. There was slightly higher progress specifically for biology diagram comprehension for the diagram-based condition, but this was not statistically significantly different. Our risk of bias analysis identified 'some concerns' with the potential selection of reported results due to lack of a formal pre-planning of analysis and missing outcome data. In all other areas the risk of bias was rated as 'low'. We have rated this as having 'some concerns' for risk of bias overall.

Coleman et al. (2018). This was a study into the effect of diagram design on comprehension of science texts. The researchers conducted a randomised experiment with 213 fourth-grade students in six elementary schools across four U.S. states. First, the classrooms were randomised into exposure—either the water cycle topic or the circulatory system topic. Next, students were randomly assigned to one of the four conditions: text only, authentic representational text, interpretational text, and integrated text. The authentic representational text included a labelled diagram, the interpretational text also had embedded captions, and the integrated text had text surrounding the diagram with arrows connecting to the diagram. The experimenter oversaw the study over three days. Day one introduced the study and secured consent for participation. On day two, teachers were provided with reading materials, passed them out to the classes, and students completed the tasks over a period of about 30 to 45 minutes. The third day was only required because scheduling issues preventing all class interventions from being completed on day two. The researchers designed a key selection task for the outcome measures based on previous research and a reading comprehension test aligned to the curriculum. They conducted a reliability analysis which indicated adequate reliability and difficulty of the measures.

Key findings. Results indicate that, overall, the inclusion of diagrams in a scientific text had minimal or no impact in facilitating fourth-grade readers' comprehension during an independent-reading task. This finding was consistent across two text exposures on different science topics and two types of outcomes measures: lower-level term selection and higher order comprehension questions. Findings also indicated that the integrated diagram may create a condition of cognitive overload for some students. Our risk of bias analysis identified 'some concerns' with potential selection of reported results due to lack of a formal pre-planning of analysis and with missing outcome data. In all other areas the risk of bias was rated as 'low'. We have rated this as having 'some concerns' for risk of bias overall.

Cromley et al. (2016) examined the effect of a cognitive science informed curriculum including teaching diagram comprehension in biology. This was an RCT with teacher-level assignment involving

9,611 seventh- and eighth-grade students of 129 teachers in the U.S. Teachers were randomly assigned to one of three groups: business-as-usual control, content-only, and cognitive-science-based. The cognitive-science-based intervention incorporated three major components (visualization exercises, case comparisons focused on highlighting key science concepts, and spaced testing in the form of daily warm-up quizzes) that were interleaved into the same base unit (that is, Holt Introduction to Matter, Cells or Inside the Restless Earth; FOSS Diversity of Life, Weather and Water, or Earth History). A fourth principle, confronting misconceptions, also informed the design.

All seventh-grade science teachers were assigned to the same condition within each school for two consecutive years, if they remained employed as science teachers at that same school. Before implementing a modified unit, cognitive science based teachers attended three paid days of summer professional development per unit they were implementing. This was coupled with providing supportive material and school year teacher discussion. Teachers in the business-as-usual control condition received neither professional development nor the modified curriculum. Instead, students attended their scheduled classes, completed only the activities included in the standard curriculum, and then completed the end-of-unit test. To measure the outcomes, six sets of three diagram-specific items were created for each curriculum and added onto the science content knowledge measure to create six unique test forms. These six test forms were then randomly given to students in the study.

> **Key findings**. The cognitive science curriculum group outperformed the content only and business as usual groups. The Cohen's d effect sizes for cognitive science versus content only across six curriculum units were d = 0.48, 0.49, 0.62, 0.52, 0.20, and 0.21. The corresponding effect sizes for the cognitive science informed curriculum compared to the control across six curriculum units were d = 0.52, 0.41, 0.55, 0.11, 0.06, and -0.13. Study 2 examined items with diagrams specifically, with the same pattern of results. The intervention was more successful in classrooms where the teacher was teaching with the interventions for the second time, suggesting some practice in implementing the diagrammatic interventions is useful. Our risk of bias analysis identified 'some concerns' with the potential selection of reported results due to lack of a formal pre-planning of analysis. In all other areas the risk of bias was rated as 'low'.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section we focus on the study outcomes, summarised in Table B6.4. Studies identified as high relevance and quality have been marked with an asterisk.

Table B6.4: Diagrams—summary of evidence

Study	Focus	Population	Finding
		High Priority Studies	
*Bergey *et al.* (2015)	Effects of diagrams versus text on spaced restudy on biology knowledge and comprehension	• *N* = 128 • *M* age = 14.9 yrs • 1 high school, 15 classes • US	***Neutral*** • Equal, significant, and moderate- to large-sized progress in both treatment and control conditions on biology knowledge (d = .46–.51), biology diagram comprehension (near transfer; d = .31–.63), and geology diagram comprehension (far transfer; d = .59–.67). Whether delivered in a text-based format or a diagram-based format, restudy warm-ups were associated with significant growth in biology knowledge. • Slightly higher progress for biology diagram comprehension for treatment but not significant.
*Coleman *et al.* (2018)	Effect of diagram design on	• *N* = 213 • 4th grade	***Neutral for younger children*** • Results indicate that, overall, the inclusion of diagrams in a scientific text had minimal or no impact in facilitating fourth-grade readers' comprehension during an independent-reading task (η^2 for conditions

	comprehension of science texts	• 6 elementary schools across 4 states • US	= .003 and .09). This finding was consistent across two text exposures on different science topics and with two types of outcomes measures—both lower-level term selection and higher order comprehension questions. Findings also indicated that the integrated diagram may create a condition of cognitive overload for some students.
*Cromley *et al*. (2016) (Study 2)^	Effect of cognitive science informed curriculum including teaching diagram comprehension in biology (Study 2)	• N = 9,611 7th and 8th grade students, 129 teachers. • US	***Positive*** • Study 2 examined items with diagrams specifically. With the cognitive science condition as a baseline (so negative effect indicating positive cognitive science effect), diagrams group outperformed the control ($b_{control}$ = –0.495, t = –2.247, p = .027, partial η2 = .05) and the content only group ($b_{content}$ = –0.642, t = –2.753, p = .007, partial η2 = .07) • The intervention was more successful in classrooms where the teacher was teaching with our interventions for the second time, suggesting some practice in implementing the diagrammatic interventions is useful.
Larger Studies (pupil n > 500) (Medium Priority)			
There were no larger studies at the medium priority level			
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
Booth *et al*. (2012)	Effect of diagrams, stories and equations on algebra problem solving	• *N* = 373 • Aged 12-14 yrs • 1 middle school • US	***Positive for older*** ***Negative for younger*** • Overall, diagrams are beneficial additions to story problems for more accomplished students. Unfortunately, results from both experiments also suggest that younger middle school students, and especially those with low math ability, do not benefit from the diagrams. Error analysis suggests that the main barrier to successful diagram use in sixth grade was the inability to extract a correct conceptual understanding of the problem from the diagram.
Butcher and Aleven (2013)^	Effect of rule-diagram mapping on geometry learning	3 experiments. All from 1 school in the US, all 10th grade Expt.1: • *N* = 96 Expt.2: • *N* = 109 Expt.3: • *N* = 83	***Positive for student generated*** • Connecting diagram elements to domain rules via student-generated highlights supported long-term learning about these rules (Experiment 3), making these same connections by interacting with solved quantities was ineffective (Experiment 1). Interacting directly with diagrams appeared to facilitate spontaneous processing of rule–diagram mappings, but providing visual representations of rule–diagram mappings negated the effects of interaction (Experiment 2). Providing visual representations of rule–diagram mappings was not as effective as scaffolding student generation of these mappings (Experiment 3).
Cromley *et al*. (2013b)	Effect of teaching diagram comprehension on comprehension of biology diagrams	• *N* = 143 • 9th grade • 12 classes • US	*Three diagram conditions. Engagement and verbalisation helps.* • 1) Self-explanation in diagrams, and 2) Student-Completed Diagrams Verbal outperformed 3) student-Completed Diagrams Visual. The latter seemed not to have demanded enough of the students to lead to efficient learning.
Kolloffel *et al*. (2009)	Effect of representational format on maths learning from an interactive computer simulation	• *N* = 123 • *M* age = 15.6 yrs • The Netherlands	***Negative for diagrams*** • Mean scores and SD of the groups were as follows (highest to lowest): ○ Text and arithmetic (30.9, 4.3) ○ Text (29.3, 5.0) ○ Arithmetic (28.2, 5.8) ○ Diagram and arithmetic (27.9, 4,1) ○ Diagram (26.5, 4.9) • Cognitive load measures suggested that diagrams help to reduce extraneous cognitive load in complex domains.
Purnell *et al*. (1992)	Effect of technical	4 experiments, all from 1 high	***Neutral***

	illustrations on cognitive load and learning in geography	school in Australia Expt.1/2/3/4 *N* = 44/29/52/100 boys *M* age = 16.2/14.7/ 13.6/16.9 yrs	• Overall, the comparison between combined and split attention approached but failed to reach • statistical significance [F(1,43)= 3.49, ,05<p<0.1]. However, one of the interactions between attention and trials did reach significance [F(1,43)=58.29, p<.01], This suggests that the difference in performance by students between the combined and split conditions was statistically significant for one or more of the three trials.
Swanson *et al*. (2013)	Effect of visual and schematic cognitive strategies on mathematics problem solving of children at risk of maths difficulties	• *N* = 120 • 3rd grade • US	***Positive for students with maths difficulties*** • When compared to the control condition, an advantage was found on post-test problem solving and calculation accuracy for children with MD for the visual-schematic strategy. However, all strategy conditions facilitated similar post-test performance in correctly identifying problem solving components relative to the control condition. • Use of diagrams supported mapping of the numbers from the text for a direct translation into a set of computations. In addition, the visual-schematic strategy may have provided a technique that allowed focus on the relevant aspects of the task without being distracted by irrelevant information.
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Carson *et al*. (2003)	Effect of diagrammatic and text-based instructions on chemistry learning	2 experiments • 1 secondary school • Australia Expt.1: • *N* = 24 • Aged 12-13 yrs Expt.2: • *N* = 28 • Aged 13-14 yrs	***Positive for high complexity*** ***Neutral for low complexity*** • Experiment 1 – as the intellectual complexity increased between Tasks 1 and 2, students benefited from the diagrammatic representation of Task 2. There were no performance differences between a text and a diagrammatic format for the low element interactive Task 1. • The findings of Experiment 2 replicated those of the first experiment, with a diagrammatic format resulting in superior learning to an equivalent text-based format but only for those aspects of the task that were high in element interactivity. There was no advantage of diagrams for those aspects of the task low in element interactivity.
Chen *et al*. (2019)	Effect of diagramming and summarising on learning from scientific texts	• *N* = 73 • Aged 13-14 yrs • 1 school • China	***Positive for student generated*** • Drawing diagrams was more effective than writing summaries in grade 7 (d = 0.19) and grade 8 (d = 0.32) as it facilitates the representation of more of the important details from the material being learned.
Chu *et al*. (2017)	Effect of diagrams on algebra equation problem solving	• *N* = 61 • Aged 12-13 yrs • 1 school, 4 classes • US	***Positive*** • The presence of diagrams increased equation-solving accuracy (d = 0.45) and the use of informal strategies (1.00). This diagram benefit was independent of student ability and item complexity. The benefits of diagrams found previously for story problems generalized to symbolic problems. The findings are consistent with cognitive models of problem solving and suggest that diagrams may be a useful additional representation of symbolic problems.

Cromley *et al.* (2013a)	Effect of teaching diagram comprehension on comprehension of biology diagrams	• *N* = 61 • *M* age = 15.5 yrs • 1 school • US	***Positive*** • When implemented with modest fidelity, COD was associated with statistically significantly better student gains in comprehension of biology diagrams (raw score increase of 30% across literal and inferential items), as compared to BAU control • studied in two classrooms taught by the same teacher. Analyses of student workbook entries also showed that higher gains from pre- to post-test on the biology diagrams measure were associated with more inferential activity and less verbatim copying from the textbook.
Reisslein *et al.* (2015)	Effect of colour-coded diagrams on learning about electrical circuits	• *N* = 74 • *M* age = 14.6 yrs • 1 high school • US	***Two diagram conditions – suggests colour more effective*** • An ANCOVA was conducted using the total post-test score as dependent variable, experimental condition as independent variable, and instructional time as covariate. The results revealed that the colour group significantly outperformed the black and white group on the post-test measure (d = 0.56, 95 % CI = 0.10, 1.03)

* High priority study identified for in-depth analysis; ^ = study included for more than one strategy.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in the table below.

Table B6.5: Diagrams—quality of evidence assessment (based on the GRADE approach)

Strategy	Diagrams
Number of studies	There are 14 studies in this area of which three were rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	All studies are randomised experiments.
Risk of bias	Our risk of bias assessments on the high-quality papers identified some concerns with missing outcome data (attrition) for two of the three studies and with the potential selection of reported results for all three studies. We rated all studies as having 'some concerns' with risk of bias. However—as per other sections—our assessment is that lack of a pre-planned analysis without any reported further issues relating to the selection of results (for example, a post-hoc 'dredging' for positive results) might be considered low risk. We judge, therefore, there to be at least one strong study in this area.
Inconsistency	**Result consistency**. There were nine positive results, three neutral or mixed results, and two negative results, one of which was a sub-experiment and specifically a negative result for younger (sixth grade rather than eighth grade) children.
Indirectness	**Practice heterogeneity**. As with the previous section, there was some practice heterogeneity relating to the conditions being compared. All studies assessed the additional effect of diagrams, but there were variations in how the diagrams were engaged with, generated, their content, and their role in the context of other learning materials and information. **Population, measure, and outcome heterogeneity**. Twelve out of the 14 studies in this area were for students between the age of 12 and 16 and 13 out of 14 were in maths and/or science. In this sense, the sample is relatively narrow. There was some variation in pupils in focus, with one study specifically focused on students with maths difficulties and several others containing mixed ability groups and finding differences in this respect, an issue we return to below. **Design and delivery**. Ecological validity was relatively low with most of the studies (11) delivered via workbooks or computer programmes with minimal teacher input. There were three studies where teachers received professional development and resources (such as workbooks) and then implemented these in their classroom.

Imprecision	**Group sizes**. Of the 14 studies, seven were small (N < 100), six were small-moderate (101 < N < 400), and one was large (N = 9,611). Overall, these results will have low precision. We judge Cromley et al. (2016, Study 2)^ to provide the highest precision estimates of impact. Results were as follows: • Cohen's d effect sizes for cogsci vs. content only across six curriculum units: 0.48, 0.49, 0.62, 0.52, 0.20, 0.21; and • Cohen's d effect sizes for cogsci vs. control across six curriculum units: 0.52, 0.41, 0.55, 0.11, 0.06, -0.13.
Publication bias	There is insufficient or no evidence of publication bias.
Other considerations	None.
Overall confidence	**Low (++)** Our confidence in the effect estimate is limited: the true effect may be substantially different from our estimate.
Confidence reasons	Key reasons for downgrading of certainty in this area: - limited sample: 12 out of the 14 studies in this area were for students between the age of 12 and 16 and 13 out of 14 were in maths and/or science; - some inconsistency in results; - most studies were small or in the lower range of our medium-sized range; the largest four studies provided mixed evidence; and - ecological validity was relatively low, with most of the studies (11) delivered via workbooks or computer programmes, with minimal teacher input.

Summary of findings for this strategy

Main finding. As with the findings relating to visual aids, the evidence suggests that diagrams when learning are mostly helpful but frequently have no effect and can sometimes be harmful. This mixed picture is made more positive (but arguably truer to the theory) when qualified. Specifically, this evidence suggests that diagrams tend to be helpful for older children to support learning of more complex content in maths and science.

Estimated impact. The highest precision estimates we have (Cromley et al., 2016, Study 2) suggest that moderate effect sizes are possible but, with the variation in outcomes in this study, a limited range of studies, the lack of studies providing precise effect estimates, and some reporting negative effects, we are not able to confidently estimate an effect range.

Confidence in impact estimate. Our confidence in this result is low due to issues with the limited sample, inconsistency, and low ecological validity in the group.

Heterogeneity. There was one study with a positive impact for students at risk of maths difficulties; there, the diagram had helped students focus on relevant aspects of the problem. However, negative results suggest that diagrams can also increase cognitive load, to the detriment of learning. We return to discuss this result and start to tease apart some of these factors in the discussion and questions section.

Strategy 12: Spatial, visualisation, and simulation approaches

Concise definition

Spatial, visualisation, and simulation approaches support children to imagine content or representations of it, often in order to simulate, manipulate, or organise concepts and schemas across time or space.

Full definition and description

Spatial, visualisation, and simulation approaches support children to imagine content or representations of it, often in order to simulate, manipulate, or organise concepts and schemas across time or space. In some cases, such as in geometry, visualisation is inherent to the learning objective, in others, it is used as a form of retrieval and rehearsal (for example, imagining a story), and in other cases, it is used as a scaffold for problem solving (that is, visualising as a way of analysing or anchoring a learning object in memory).

Selected examples

Examples of this strategy from our database include:

- Hawes et al. (2017) provided young children with a series of lessons to support spatial visualisation. These included arranging 2D square tiles and cubes into configurations, a symmetry game with shapes, a lesson where children predicted how many tiles were needed to cover a mat, and a 'garden patio' creation activity.
- Bokosmaty et al. (2017) examined learners' use of software to manipulate geometric shapes, learners observing a teacher doing to same, and a static condition with fixed shapes (for example, triangles with various properties).
- In De Koning et al. (2017), children were supported to mentally simulate a story in a text. They were encouraged to pay attention to sensory information over eight 30-minute sessions over four weeks. The programme was presented to children as a 'movie director training' in which they create a movie from the text in their head. Scaffolding techniques to support students' imaginative processes were gradually 'faded' out as the children become more confident.
- Barner et al. (2016, 2018) taught children how to use the mental image of an abacus and use this as part of calculation tasks in maths.

Evidence for this approach

There were seven studies focused on spatial, visualisation, and simulation approaches. Of these, two were graded as high priority Full details of all medium and high studies are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed for this strategy are characterised as follows:

- **Pupil age and characteristics** Studies in this section focused on primary-age students—children from 4 to 12 years old were represented. There was a good spread across the primary age range within the studies.
- **Location**. Given the small number of studies, there were a wide range of countries represented in the data. There were studies from the U.S. (two), India (one), Australia (two), the Netherlands (one), and the U.K. (one).

- **Learning areas**. Studies in this area were, with only one exception, focused on spatial visualisation on mathematics outcomes, including work in areas that have an inherent spatial aspect (geometry) and those that ostensibly do not such as number. There was one study of mental simulation on reading comprehension.
- **Outcome Measures**. The outcome measures used within this section were strong relative to other sections. Four studies combined researcher-designed tests aligned to content with standardised measures (such as national tests or tests with known psychometric properties). One study used a standardised curriculum test; two used researcher designed tests aligned to the content.
- **Design and delivery**. Five of the seven studies were designed and delivered by researchers. There were two studies delivered by regular teachers (who received professional development training on the strategy). Two studies were relatively short, consisting of one and two experimental sessions only, respectively. Two studies were of a moderate duration of three to four weeks. Three studies were longer, ranging from 32 weeks to three years.

High priority studies in this area

There were two studies in the spatial, visualisation, and simulation strategy category that were rated as having high strength and validity of evidence. We conducted in-depth analysis of these studies and have completed a full risk of bias assessment, summarised in the appendix.

Barner et al. (2016). This study examined the effect of mental abacus instruction on maths achievement. Barner and colleagues have also published a second study in this area (Barner et al., 2018), included in this section. We have selected the larger and longer study for risk of bias assessment. This study was a randomised controlled trial with an assignment at classroom level for 204 primary school children, aged five to seven, in three classes in India. Treatment group students received an extra three hours per week of mental abacus instruction. Control group students received and extra three hours per week of supplementary maths tuition (non-abacus). Both groups studied the school's standard (non-abacus) mathematics curriculum. The mental abacus training was delivered by an experienced mental abacus teacher (not the children's regular teacher). The study spanned three years. Outcome measures combined several standardised tests with in-house assessment of place value and arithmetic. The standardised tests of learning outcomes were a subtest of the Woodcock–Johnson Tests of Achievement (WJ–IIIC) and the Math Fluency subtest of the Wechsler Individual Achievement Test (WIAT–III).

Key findings. The results revealed that the mental abacus group outperformed the control group on three of the four mathematics tasks, with statistically significant effect sizes of d = 0.60 for arithmetic, 0.24 for WJ–IIIC, and 0.28 for place value. The WIAT-III outcome showed a small, positive but statistically insignificant effect (d = 0.13). The improvement was mediated by children's individual visual working memory capacities at the beginning of the study (higher starting WM predicted greater improvement). In our risk of bias assessment there were 'some concerns' with the randomisation process, and (pre-planning) of the reported results. Overall, we rated this study as 'some concerns' for risk of bias.

Lowrie et al. (2019). This study examined the effect of spatial visualisation training on maths performance. A quasi-experimental design was used. Ethical restrictions set by governing educational jurisdictions prevented random assignment. Pre-test comparisons were conducted to assess group equivalence, with differences in two pre-tests found not statistically significant. The study involved 327 students aged 10 to 12 from 17 classes in ten schools in Australia. The treatment group received a diverse range of spatial visualization activities and topics (such as reflection, symmetry, and paper-folding) used in lieu of geometry lessons. The control group received their standard mathematics

instruction (a 'business as usual' condition). This was a three-week intervention consisting of 60-minute sessions twice a week. The intervention was delivered by teachers who received ten hours of CPD on spatial visualisation. The outcome measures were a standardised measure of spatial reasoning (SRI) and a mathematics test developed using items from Australia's National Assessment Program (NAPLAN) Numeracy test.

Key findings. The results saw the spatial visualisation group outperforming the control group on spatial reasoning performance (d = 0.40) and mathematics performance (d = 0.39). In our risk of bias assessment this study had a 'high' risk of bias due to the non-random assignment (although, as noted above, this does not appear to have resulted in observable group imbalance) and 'some concerns' relating to the selection of reported results. Overall, due to the non-randomisation of conditions, this study was rated as having a 'high' risk of bias.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section we focus on the study outcomes, summarised in Table B6.6. Studies identified as high relevance and quality have been marked with an asterisk.

Table B6.6: Spatial, visualisation, and simulation approaches—summary of evidence

Study	Focus	Population	Findings
High Priority Studies			
*Barner *et al*. (2016)	Effect of mental abacus instruction on maths achievement	*N* = 204 Ages 5-7 1 primary school, 3 classes India	***Positive*** • Mental abacus group performed better. The results revealed that the mental abacus group outperformed the control group on three of the four mathematics tasks, with statistically significant effect sizes of d=.60 (95 % CI =0.30, 0.89) for arithmetic; 0.24 (95 % CI =-0.05, 0.52) for WJ–IIIC; and d = 0.28 (95 % CI = 0.00, 0.57) for place value. The WIAT-III outcome showed a small, positive but statistically insignificant effect (d = 0.13, 95 % CI = -0.15, 0.42).
*Lowrie *et al*. (2019)	Effect of spatial visualisation training on maths performance	*N* = 327 Ages 10-12 10 schools, 17 classes Australia	***Positive*** • Spatial visualisation group improved on spatial reasoning performance (d = .40, t(17) = 4.59, p = .04) and mathematics performance (d = .39, t(17) = 6.95, p = .016) compared to control group.
Larger Studies (pupil n > 500) (Medium Priority)			
There were no larger studies at the medium priority level			
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
Barner *et al*. (2018)	Effect of mental abacus instruction on maths achievement	*N* = 164 1st and 2nd grade (Ages 6-8) 1 primary school, 24 classes US	***Neutral*** • No evidence of significant differences between groups on any of the mathematical achievement measures. Also, no difference in executive functioning.
De Koning *et al*. (2017)	Effect of mental simulation on reading comprehension	*N* = 143 Ages 10-11 5 primary schools The Netherlands	***Positive*** • Grade 3: Compared to the control group, children who received the mental simulation training showed improved performance on general reading comprehension • Did not affect scores in Grade 4

Gilligan *et al.* (2019)	Effect of spatial training and instruction type on maths performance	*N* = 250 Ages 8-9 (Year 3) 6 primary schools UK	***Mixed, but generally positive*** • Mental rotation and spatial scaling training led to significant gains in mental rotation, and spatial scaling respectively ('near transfer') • Selective effects for more classroom-relevant outcomes: ○ Mental rotation training: improved missing term problems ('far transfer') ○ Spatial scaling training: improved number line estimation ('far transfer') ○ No effect on geometry task scores between groups ○ Implicit compared to explicit instruction generally made no difference
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Bokosmaty *et al.* (2017)	Effect of manipulating shapes (using mouse movements) on geometry problem-solving	*N* = 60 Ages 9-11 1 school Australia	***Positive*** • 'Similar' items: Manipulation condition significantly outperformed the conventional learning condition (d = 1.58, 95 % CI = 0.87, 2.29) and observing manipulation conditions (d = 1.00, 95 % CI = 0.34, 1.65) • 'Transfer' items: Both manipulation (d = 1.38, 95 % CI = 0.69, 2.07) and observing manipulation (d = 0.84, 95 % CI = 0.20, 1.49) conditions outperformed the conventional condition • Both manipulation conditions had lower reported cognitive load than control
Hawes *et al.* (2017)	Effect of spatial visualisation training on maths performance	*N* = 65 Ages 4-7 3 elementary schools, 12 classes US	***Mixed, but generally positive*** Some benefits: • Significant benefit of spatial visualisation on visual-spatial geometry task ('near transfer') • Significant benefit on symbolic comparison (ηp2=0.10), but not non-symbolic comparison, or number knowledge ('far transfer' tasks)

* High priority study identified for in-depth analysis.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in the table below.

Table B6.7: Spatial, visualisation, and simulation approaches—quality of evidence assessment (based on the GRADE approach)

Strategy	Spatial, visualisation and simulation approaches
Number of studies	There are seven studies in this area of which two were rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	Six of the seven studies are randomised experiments. One was a quasi-experiment with non-random assignment to conditions.
Risk of bias	Our risk of bias assessments on the high-quality papers raised 'some concerns' and 'high' concerns with the randomisation process of the papers and 'some concerns' with the potential selection of reported results for both. One was rated as having 'some concerns' overall and one 'high' risk of bias. Therefore, we cannot be confident that there are any papers in this area that have low risk of bias.
Inconsistency	**Result consistency**. The results were mostly positive (4 of 7) with two with neutral or mixed results tending towards positive and one neutral result.

Indirectness	**Practice heterogeneity**. All studies in this area, with one exception, were focused on the effect of spatial visualisation in maths, including geometry and number. Even within this group, we are unsure whether curriculum areas with an inherently spatial aspect (geometry) can reliably be grouped with studies of number with spatial elements or concepts such as number lines and abacuses. **Population, measure, and outcome heterogeneity**. The student population for these studies spanned ages 4 to 12 and therefore represents the primary but not the secondary age range. There were several standardised measures used of similar outcomes. **Design and delivery**. Five of the seven studies were designed and delivered by researchers. There were two studies delivered by regular teachers (who received professional development training on the strategy).
Imprecision	**Group sizes**. There were two small studies (N < 100) and five small to moderate sized studies (101 < N < 400) in this area. There were only seven studies in total, and no larger studies. Overall, and with the high priority studies particularly in mind, the results suggest a small to moderate effect (d = 0.1–0.5). High priority study results were as follows: - *Barner et al. (2016) estimate effects of d = 0.60 (95% CI: 0.30, 0.89) for arithmetic; 0.24 (95% CI: -0.05, 0.52) for WJ–IIIC; and d = 0.28 (95% CI: 0.00, 0.57) for place value. The WIAT-III outcome showed a small but statistically insignificant effect (d = 0.13, 95% CI: -0.15, 0.42). - *Lowrie et al. (2019): spatial visualisation group improved on spatial reasoning performance (d = 0.40) and mathematics performance (d = 0.39) compared to control group (N = 327).
Publication bias	There is no indication of publication bias.
Other considerations	
Overall confidence	**Very Low (+)** We have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of effect.
Confidence reasons	Key reasons for downgrading confidence in this group are: - this is a small strategy group of only seven studies, encompassing several related but varied approaches; - the student population for these studies spanned ages 4 to 12 and therefore represents the primary, but not the secondary, age range; and - five of the seven studies were designed and delivered by researchers.

Summary of findings for this strategy

Main finding. Overall, this area shows some promise but the evidence is insufficient to judge the effectiveness of strategies in this area, either for primary maths or more widely.

Estimated impact. Overall, these results suggest small to moderate effects for using spatial, visualisation, and simulation approaches in maths for primary school age children.

Confidence in impact estimate. We have, however, rated our confidence in this judgement as being 'very low'. As we describe in the GRADE analysis, this is a small evidence-base for which the high priority studies had 'some' and 'high' risks of bias.

Heterogeneity. We judge there to be conceptual as well as practical differences in the strategies being tested, but have not had sufficient evidence to examine these differences.

Cognitive theory of multimedia learning—overall evidence summary and conclusions

Summary of results

In this section, we have reviewed 55 (34 + 14 + 7) studies focused on the presentation of information in multiple modes. We identified three strategies for which we potentially had sufficient evidence to assess effectiveness. Our results for these are summarised in Table B6.8.

Table B6.8: Strategies related to the cognitive theory of multimedia learning—summary results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Visual representation and illustration	Thirty-four, of which three were graded as high priority.[1]	The evidence suggests that visual aids are most helpful during learning but frequently have no effect and can sometimes be harmful.	The age range of students was 7 to 18, with a good spread within this range. Note that there is, therefore, no evidence for children younger than this. Studies represented a range of subjects. Although over two-thirds were of maths and science.	Low (++)
Diagrams	Fourteen, of which three were graded as high priority.[1]	The evidence suggests that diagrams for secondary maths and science learning are mostly helpful but frequently have no effect and are often harmful.	Twelve out of the 14 studies in this area were for students between the age of 12 and 16. Also, 13 out of 14 were in maths and/or science. In this sense, the sample is relatively narrow.	Low (++)
Spatial, visualisation, and simulation approaches	Seven, of which two were graded as high priority.[1]	Overall, this area shows some promise but the evidence is insufficient to judge the effectiveness of strategies in this area, either for primary maths or more widely.	All studies in this area, with one exception, were focused on the effect of spatial visualisation in maths, including geometry and number. The student population for these studies spanned ages 4 to 12 and therefore represents the primary, but not the secondary, age range.	Very Low (+)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Conclusions about strategies in this area

Dual coding and multimedia learning

Evidence about how teachers use visual information and combine modes of information has high potential relevance across the U.K. education system for all learners and subjects. The simple changes, for example, of adding or taking away images from instructional materials, replacing written text on slides with just images, or providing diagrams, could have far-reaching implications.

In terms of the evidence we have here, however, firm conclusions have been challenging. For example, for visual aids and diagrams, when we crudely compare conditions with and without these, there are mixed results.

As we have touched on above, however, a slightly more nuanced interpretation of the theory would hold that the impact of images would depend on their decorative or informational content, their role and centrality within the learning, the format and content of other modes of information and how complementary these were, how the image was engaged with (including student generation), the student prior knowledge, the overall cognitive load, and more.

The evidence is not sufficient for us to make these distinctions and reach robust judgements on the effect sizes for subgroups and their impact on different learning outcomes and populations. As a result, we return to this question and start to tease apart some of these factors in the discussion and questions section below.

At the outset of this section, we noted that we also located studies that compared the effect of images with audio or animations on learning. The former of these is arguably more relevant to dual coding theory than some of the studies above that combine written text (visual) with images (visual), although as we discuss below, drawing on the cognitive theory of multimedia learning. Elsewhere, the simple equation of information presentation types with working memory processing of this information is complex.

Overall, it has been disappointing that in an area of evidence where we originally identified 122 studies that there are so few clear and robust tests of the theoretical principles in applied settings.

Evidence-informed discussion and questions

Principles and moderating factors

Cognitive theory of multimedia learning and dual coding theory

How do the cognitive theory of multimedia learning and dual coding relate?

At the outset of this review area, we described how we had framed this section in terms of the Cognitive Theory Of Multimedia Learning (CTML) rather than focusing on dual coding, as per our original intention. Our decision involved a trade-off between selecting a narrower theory—dual coding—and having a clearer focus for our analysis versus selecting a broader theory—CTML—and being able to situate studies into a broader framework that can make sense of the multiple principles for the educational use of multimedia represented in the database. The CTML, like dual coding, connects to ideas from cognitive load theory but also connects to ideas related to generative learning examined in the last review area. As we expand on below, there are also links to our next review area: embodied learning. Rather than shy away from these connections, we are of the view that cognitive science strategies and concepts are neither discrete nor linear, that making connections is valuable, and that this is the section to attempt to do so.

Recall from the introduction to this area the following three assumptions that underpin the CTML:

> 1. *Dual channels – Humans possess separate channels for processing visual and auditory information.*
> 2. *Limited capacity – Humans are limited in the amount of information that can be processed in each channel at one time.*
> 3. *Active processing – Humans engage in active learning by attending to relevant incoming information, organising selected information into coherent mental representations, and integrating mental representations with another knowledge.*
>
> *(Mayer, 2021, p.34)*

The first of these assumptions is central to ideas relating to dual coding. As we described, the two dual channels—as set out in the Baddeley and Hitch (1974) model of working memory—are as follows:

- **a visuospatial sketchpad** that deals with visual and spatial information: here we process images 'synchronously', seeing them all at once, their links and location in space; and
- **a phonological loop** that deals with auditory information: here we process information 'sequentially', experiencing and playing auditory information back to ourselves in a 'loop'.

The second assumption connects to cognitive load theory (as per our review area on managing cognitive load). The third assumption connects to our section on Working with Schemas, which also included discussion of generative learning. Mayer (2021, p.43) describes active processing as the need to select (that is, attend to) and organise words and images in working memory, make connections between them and integrate them 'with relevant prior knowledge activated from long-term memory'.

What makes the Cognitive Theory of Multimedia Learning a cognitive theory of learning as opposed to a more general theory of learning? As Mayer (2021, pp.3, 10) explains, multimedia can be understood in terms of the media used to convey information (for example, a speaker and computer screen), as a mode of presentation (for example, words and pictures), or as sensory modalities (for example, auditory and visual). The latter connection makes CTML a cognitive theory of learning, connecting as it does with the 'dual channel' account of the working memory outlined above. It is this dual channel theory that remains at the core of our focus. Below, we consider the prevailing basic scientific account of the educational implications of dual coding theory and connect this to some of our wider evidence and practitioner perspectives. We then examine our wider evidence and teacher perspectives on (a) visual aids, (b) animations, and (c) multimedia learning and—as we do—we slowly expand our frame of reference from the core of dual coding theory to the broader CTML.

Why might dual coding of information be beneficial for learning?

In Caviglioli (2019, p.20–21), Paul Kirschner describes two main benefits of dual coding. First, he notes that dual coding learning allows the learner to 'benefit from access to both visual and verbal memory capacity'. Mayer (2021, p.7) describes this benefit as the 'quantitative explanation'. Second, that coding information produces two information 'traces' which, according to Kirschner, (a) will be 'stronger than one single trace' and (b) 'allows you to remember or recognise the information in two different ways' (Caviglioli, 2019, p.21). Again, Mayer (2021) agrees: he views the quantitative explanation as 'incomplete' and describes a 'qualitative explanation' for the benefits of dual coding as follows:

> *'The qualitative rationale is that words and pictures, while qualitatively different, can complement one another and that human understanding occurs when learners are able to mentally integrate corresponding pictorial and verbal representations.'*
>
> *(Mayer, 2021, p.7)*

Therefore, dual coding is thought to be beneficial because it makes better use of working memory capacity while offering complementary forms of information that promote both encoding, and retrieval.

What are the potential wider benefits of multimedia learning?

Dual coding, and its use of imagery and audio, is frequently bound up with wider multimedia learning principles in both scientific, popular, and professional accounts we reviewed. When we examine teacher descriptions of dual coding in our practice review interviews and questionnaires, there are a range of practices and principles that teachers link to dual coding, which tap into the core ideas

discussed above while often going beyond. There is a grey area between cognitive theories of learning that we can link back to the basic scientific understanding of the brain and learning and more general educational or technology-based learning principles.

Other cognitive scientific principles that are relevant include the following ideas summarised in Caviglioli (2019):

- First, that we know **non-verbal information is processed synchronously, while verbal information is processed sequentially**. This means that auditory information is '**transitory'** and needs to be played back on a 'loop' to retain it in working memory (like verbally rehearsing a telephone number in your head to help you remember it). A teacher might replace text with images and then provide an oral explanation—this would make use of both channels, but potentially run up against students' (in)ability to retain large amounts of transitory auditory information.
- A second idea in this area is the so-called '**visual argument**' (Caviglioli, 2019, p.26, and see Vekiri, 2002)—that it is easier (more '**cognitively efficient**') to process information visually, as we can often search, recognise, and see connections in images more readily than in equivalent texts.
- Third, a link is often made between dual coding of information and **embodied cognition**. Embodied cognition relates to the role of the body in forming concepts and supporting cognition through, for example, enacting concepts, gesture, and movements. Embodied cognition is the focus of the next area we review.

We do not claim this to be an exhaustive list of related concepts but rather a summary of the most prominent ideas in practitioner-focused accounts of the basic science we have consulted. An implication of the varied principles and concepts discussed above is that the success of multimedia presentation of information is likely to be affected by multiple, sometimes countervailing, cognitive mechanisms. It also suggests the need for careful attention to which factors might be operating in a given teaching and learning context or study.

What does our wider evidence suggest about the efficacy of dual coding, and its principles?

Within our wider evidence-base, we identified several studies that looked specifically at the combination of audio and visual information compared to a single channel presentation. Although we judged there to be an insufficient weight of evidence (that is, number of high and medium priority studies) and many were not designed to test the core principle, these studies nonetheless provide important indicative evidence. We summarise these studies (all medium priority)below:

- **Harskamp et al. (2007)** examined the modality principle (the idea that students learn better from graphics and spoken text than from graphics and printed text) applied to secondary school students' learning of biological concepts with web-based, multimedia lessons in their school. For some students, the science lessons contained a series of illustrations with concurrent narration. In contrast, for other students, the science lessons contained a series of illustrations with concurrent on-screen text. They found that the modality principle was strong for learner understanding (transfer) but not for retention, and that the effect was strong for 'fast' but not 'slow' learners.
- **Scheiter et al. (2014)** explored the effects of different multimedia designs for learners' reading comprehension and scientific literacy (biology). Students (with an average age of 15.1 years) learned about cell reproduction via different types of media (text only versus text plus animations)

and text modality (spoken versus written versus spoken and written). They found that adding animations to text and using spoken rather than written text improved only immediate recall. However, for delayed recall, a multimedia effect was observed for learners with higher levels of scientific literacy. They also found that a redundant presentation of text proved harmful, especially on delayed performance measures.

- In two studies, **Leahy et al. (2003)** and **Leahy and Sweller (2016)** looked at the effectiveness of audio-visual based instruction.
    - **Leahy et al. (2003)** compared visual plus audio presentations with visual-only presentations. Their first experiment found that the former was superior as neither the auditory nor the visual material could be understood in isolation. However, in their second experiment they found that when non-essential explanatory text was presented audially with similar written text contained in a diagram, it hindered learning because it created a redundancy effect. Therefore, they concluded that the effectiveness of multimedia instruction depends on how and when auditory information is used.
    - **Leahy and Sweller (2016)** similarly explored the length and complexity of auditory and visual text instructions. They found that shorter, audio-visual information was better than visual-only information, but longer, audio-visual information was worse than visual-only information.
- **Wong et al. (2012)** similarly looked at the length of segments to be learned and the effect of (a) animations as opposed to static graphics and (b) audio-visual information instead of visual information only. Again, findings supported their hypothesis that animations would be superior to static graphics for transient information presented in short sections but not for transient information in long sections.
- **Lee and Mayer (2015)** discuss the impact of audio versus audio plus video on learners studying material in their second language (Korean students learning about Antarctica in English). They found that when the audio was in English, the audio plus video group scored significantly or marginally higher than the audio group on a subsequent comprehension test. However, with a second experiment *with university students*, they found that when the audio was in Korean, comprehension scores of college students did not benefit from the added video.
- Exploring the impact of prior knowledge on the use of multimedia presentations, **Leslie et al. (2012)** conducted two experiments using the science topics of magnetism and light with Year 5 students with no prior knowledge of the topics and Year 6 students who had studied the topics previously. Results indicated that the older students with prior knowledge of the topic learned more from the auditory-only presentation. For these students, the addition of visual information was redundant and thus they were disadvantaged by the use of an audio-visual presentation. However, for younger students with no prior knowledge of the topic, the difference between mean scores reversed. They therefore concluded that some of the younger students might require a visual presentation to make sense of the auditory explanation.

While we are not in a position to systematically review applied dual coding theory from such results, we note that there are several principles at play: learner prior knowledge and ability to process information, transfer versus retention of learning, immediate versus delayed retention, the potential redundancy of information, and the link between transient information and presentation length. These studies, therefore, do not suggest that the successful application of dual coding is straightforward; teachers are likely to need to consider and balance multiple principles (see below for more discussion of this point). Another complexity that we have encountered in accounts of dual coding is the distinction made by Mayer between the mode of presentation (words or pictures in curriculum resources), the sensory memory mode (eyes or ears) and the mode of representation in

working memory. As Mayer (2021, p.41) explains, visual images can be mentally converted to audio information and vice versa. A picture of a cat might trigger you to mentally hear the word 'cat'; the written word 'cat' might similarly produce audio information, or perhaps the visual image of a cat, and so on. This—along with the transitory information effect—appears to make it challenging in practice to know the extent to which a particular multimedia presentation will produce cognitive load in the phonological loop or the visual sketchpad. In real classroom situations, effects are likely to depend on learning content as well as the learners themselves.

While these results suggest complexity in the application of dual coding theory, many teachers in our interviews and questionnaires felt that dual coding was an approach that could successfully reduce cognitive overload (both by using both channels and this allowing reduction of redundant text and spoken word) and emphasise and focus attention on key ideas and promote learning.

Variation in the practice or teaching and learning context

Practitioner-focused accounts of dual coding, such as Caviglioli (2019), describe a range of potential benefits of dual coding and the visualisation of information. These include the ability of visual information to direct attention, trigger prior knowledge, manage cognitive load, build schemas, transfer information to working memory, and motivate students (p.40). Below we have examined the perceived benefits in teacher accounts of dual coding and, where available, connected these to our wider evidence-base. We have organised this discussion into three areas focusing on, respectively:

1. visual aids;
2. animations; and
3. the use of multimedia (and principles of the CTML) more generally.

Visual aids

Are visuals more central to some areas of learning and subjects than others? Which? Why? Does the value of visual aids or visual learning depend on the decorative versus informational content of the images? What counts as decoration or information?

We remind the reader that the group of studies for visual representation and illustration was sufficient for inclusion in the main review. There we concluded as follows:

> For visual aids and diagrams, when we crudely compare conditions with and without, there are mixed results. A slightly more nuanced interpretation of the theory, however, would hold that the impact of images would depend on their decorative or informational content, their role and centrality within the learning, the format and content of other modes of information and how complementary these were, how the image was engaged with (including student generation), the student prior knowledge, the overall cognitive load, and more. The evidence is not sufficient for us to start to make these distinctions and reach robust judgements on the effect sizes for subgroups and their impact for different learning outcomes and populations.

This general picture of mixed results and complexity is consistent with the discussion immediately above about (a) the number of theoretical principles rooted in cognitive science potentially at play and, (b) the mixed results in the wider evidence and need to explore conditionality (for example, on moderating factors) in the results. Below, we report several examples of teachers describing how they made sense of dual coding. This gives a picture of different ways in which teachers have made sense of this area and how it should be applied in practice.

Visualising aids—especially with abstract or complex ideas

- 'We do sometimes a picture with definitions. There are so many grammatical skills, how do you represent them consistently with symbols? In maths more so, because you can represent problems with visual representations, but it is more difficult with reading and writing' (Interviewee 12).
- 'I find dual coding is particularly beneficial to teaching chemistry because many areas are so abstract' (questionnaire response).
- 'These principles lie at the heart of trying to replicate the experience of learning a language in the artificial environment of the classroom. The attachment of images to words to stimulate memory is central to language learning' (questionnaire response).
- 'On all lesson slides that are taught, dual coding occurs in [the] form of pictures that help give clarity on the meaning of words and more complicated concepts that are being explored' (questionnaire response).
- 'Dual coding, I think, really changed things for me. In literature obviously we spend a lot of time talking about plot and characters, and a few years ago I would have just done note taking and reading through notes and really kind of low utility revision methods, whereas now I'm kind of like "right do a picture of how that character acts, like as an image" and allowing the books to be a lot more messy!' (Interviewee 4).
- 'Generally, on our lesson plans, we use Powerpoint. We try to make sure there is an image that helps pupils' understanding. We want to avoid images just for the sake of images, it has to be useful otherwise it shouldn't be there. [Interviewer: How do you know when something is necessary or not?] I guess it is trying to look at it from a child's perspective. Giving an example: 'exuberant'. How do you try to explain what it means? But maybe seeing someone on that stage helps? It is not easy. Because the words have meaning, but you don't want to simplify that meaning, so that, for example, exuberant just becomes enthusiastic. That has similarities, but you want to make sure you have nuances' (Interviewee 11).

As a memory aid

- 'Students then have a worksheet with just the icon for each word that they must complete with the word and its definition. Repeat task until all students have mastered the knowledge' (questionnaire response).
- 'I find [dual coding] particularly effective. This gives pupils something to associate their learning with. It is also something I do in every assembly. The more obscure or thrilling the visual aid the better it helps them memorise the subject matter' (questionnaire response).

Organising and connecting

- 'We have lots of key terminology in psychology, sociology, and criminology so we try to do pictures to help their understanding. In terms of helping them to process things, we also turn things into diagrams so they process information about how things are linked together. So while we are doing a particular topic, we might put it into a diagram so they can process not only the concept itself, but also how it fits into the wider context as well' (Interviewee 2).
- 'I produce one-page diagrams for each major topic, with a mixture of pictures and words containing just the key info.'
- 'We use an adapted flow map to model for children (age 11 to 13) how to plan a non-fiction essay. It works very well as a way of making sure they use connectives and structure their points. It also improves paragraph use.'

- '[I use] dual coding to break down processes and to organise learning from difficult reading material.'

This diversity of approaches and applications was also evident in our studies. We noted in our main review of visual representation strategies that, while all studies (for inclusion in the group) tested the impact of adding an image, usually as compared to a text-only condition, there was also considerable variation in the specific strategies and the teaching and learning intentions to which they were applied. Other studies in the general area, not included in the main review, compared symbols and visual representations of number lines (Moreno and Duran, 2004), visual illustrations combined with retrieval (Jägerskog et al., 2019),[25] and the use of subtitles on learning and cognitive load (Baranowska, 2020), and studies exploring the role of imagination in combination with dual coded presentations (Tindall-Ford and Sweller, 2006). One question that we were not able to resolve, but pose here, relates to the distinction between imagining and perceiving information:

Does imagining or visualising images or other information have more, or less, value than perceiving it? Does it depend on learner knowledge of the content area?

Animations

Do animations produce higher cognitive load and distraction than static images? Can this be helpful? How does this link to spatial learning?

Animations did not feature heavily in teacher accounts of dual coding and multimedia in our practice review data collections. Nonetheless, there was a substantial body of literature in our database that looked at the impact of using animations for learning.

Some of the studies focused specifically on animations for maths and science learning:

- **Barak and Dori (2011)** investigated the impact of using animated movies and supplementary material for studying science as opposed to only textbooks and still pictures. Their results showed that animated movies promoted various thinking skills among students and enhanced scientific curiosity, language, and thinking. These findings are explained by the use of both visual-pictorial and auditory-verbal capabilities of students.
- **Dervić et al. (2019)** compared the effect of 'Physlet' animations, printed sequences of selected animation frames, versus traditionally presented static pictures when understanding about lenses. They found that the animated Physlet-based teaching generally led to higher germane load and more effective learning than the traditional approach.
- **Starbek et al. (2009)** studied the use of animations in learning genetics. Four comparable third- and fourth-grade high school student groups were taught the process of protein synthesis in a traditional lecture format by reading text, by two short computer animations, or by text supplemented with illustrations. The groups studying with animations or text supplemented with illustrations acquired better knowledge and improved comprehension skills than the other two groups. The authors thus conclude that animations and illustrations can lead to better learning outcomes when learning genetics.

[25] Their findings showed that the visuo-verbal lecture resulted in better learning than verbal presentation only, but that taking tests (retrieval practice) did not lead to better learning than restudying. They therefore conclude that it is worthwhile to use visual illustrations in teaching, but that there doesn't seem to be any synergistic effects of combining visuo-verbal presentation and retrieval practice.

- **Yang et al. (2017)** also studied the effectiveness of animations versus static pictures for learning genetics. They found that the students who learned via animation perceived less extraneous cognitive load, and achieved a better learning outcome, than those in the static pictures group. Similar to Starbek et al. (2009), they argue for the 'superiority of the animation over static picture instruction when learning micro-scientific phenomena' (p.1).
- **Scheiter et al. (2014)**, also mentioned above in relation to dual coding, explored the effects of different multimedia designs for learners' reading comprehension and scientific literacy (biology). Students (with an average age of 15.1) learned about cell reproduction via different types of media (text only versus text plus animations) and text modality (spoken versus written versus spoken and written). They found that adding animations to text and using spoken rather than written text improved only immediate recall. However, for delayed recall, a multimedia effect was observed for learners with higher levels of scientific literacy, and a redundant presentation of text which proved harmful, especially for delayed performance measures.
- **Scheiter et al. (2010)** explored the effect of augmenting worked examples with animations for teaching problem-solving skills in mathematics. Their study of 32 pupils from a German high school found that learners with hybrid animations showed superior problem-solving performance for problems of different transfer distance relative to those in the text-only condition.

A few studies also looked at the effect of animations for learning English:

- **Dindar et al. (2014)** looked at the effect of animations within the context of English language learning and drawing on cognitive load theory. They administered a computer-based English achievement test to 303 seventh-grade students, with test questions either with static graphics or with animated graphics accompanied by text. The animated graphics were found to increase the students' response time and secondary task scores, but not their test success. Also, no difference was observed in self-reported cognitive loads.
- **Fong et al. (2012)** and **Fong (2013)** looked in further detail at animations, comparing continuous and segmented animation.
    - **Fong et al. (2012)** randomly assigned 237 secondary biology students with three different trait anxiety levels into three experimental conditions: (a) text with static graphics (TSG), (b) text with animated graphics (TAG), and (c) text with segmented animation (TSG). They found that segmented animation was more effective than continuous animation and static graphics for improving learning across all levels of anxiety. Based on this, they argue that 'continuous animation does not provide sufficient time for optimal cognitive processing of information, thus inhibiting effective learning' and that 'segmented animation helps high anxiety students overcome the threat of extraneous cognitive load thus optimizing their information-processing abilities'.
    - **Fong (2013)** also looked at the effect of segmented animated graphics. Similar to the previous study, the results, this time with 171 secondary chemistry students, showed that the segmented animations were more effective than the two other conditions across all levels of spatial ability. In addition, students with low spatial ability performed significantly better with the segmented animated graphics than students in the two other groups.
- **Wong et al. (2012)**—also reported in relation to dual coding above—similarly looked at the length of segments to be learned and the effect of (a) animations as opposed to static graphics and (b) audio-visual information as opposed to visual information only. Again, findings supported their hypothesis that animations would be superior to static graphics for transient information presented in short sections but not for transient information in long sections.

These results provide indicative evidence in support of multimedia learning and touch on similar moderating principles such as transient information, cognitive load, learner prior knowledge, and so on. We must stress here that the studies briefly summarised above have not been systematically reviewed since the weight of evidence provided by these studies was not sufficient (all were medium priority, but the areas of focus and tightness to the focus cognitive science principles was not sufficient for selection).

Multimedia learning

What are the principles for combining multimedia information? Are the principles outlined by Mayer (2021) supported by evidence and relevant for school-age pupils across subjects?

We have examined our wider evidence and practitioner perspectives on dual coding, visual aids, and animations. The final consideration in this subsection on strategy variation is the broadest: multimedia learning in general. We begin by summarising key principles for Mayer's (2021) CTML followed by a brief summary of several relevant studies we located in our wider evidence.

Mayer (2021, p.53) identifies three general goals for multimedia learning design:

- reduce extraneous processing;
- manage essential processing; and
- foster generative processing.

These, of course, link back to the assumptions of the CTML (see above). Below we provide a brief summary of Mayer's 15 principles organised under these three general goals. This is necessarily brief, and we refer readers to the full text of Mayer (2021) for further details of the theory, evidence, and practice of these principles.

Box B6.1: Mayer's principles of multimedia design

Reduce extraneous processing

1. **Coherence principle**: people learn better when extraneous material is excluded rather than included.
2. **Signalling principle**: people learn better when cues are added that highlight the organisation of the essential material.
3. **Redundancy principle**: people do not learn better when printed text is added to graphics and narration; people learn better from graphics and narration than from graphics, narration, and printed text when the lesson is fast-paced.
4. **Spatial continuity principle**: people learn better when corresponding words and pictures are presented near rather than far from each other on the page or screen.
5. **Temporal contiguity principle**: people learn better when corresponding words and pictures are presented simultaneously rather than successively.

Manage essential processing

6. **Segmenting principle**: people learn better when a multimedia lesson is presented in user-paced segments rather than as a continuous unit.
7. **Pre-training principle**: people learn better from a multimedia lesson when they know the names and characteristics of the main concepts.
8. **Modality principle**: people learn better from graphics and narration than from graphics and onscreen text.

Foster generative processing

9. **Multimedia principle**: people learn better from words and pictures than from words alone.
10. **Personalisation principle**: people learn better from multimedia lessons when words are in conversational style rather than formal style.
11. **Voice principle**: people learn better when the narration in multimedia lessons is spoken in a friendly human voice rather than a machine voice.
12. **Image principle**: people do not necessarily learn better from a multimedia lesson when the speaker's image is added to the screen.
13. **Embodiment principle**: people learn more deeply from multimedia presentations when an onscreen instructor displays high embodiment rather than low embodiment.
14. **Immersion principle**: people do not necessarily learn better in 3D immersive virtual reality than with a corresponding 2D desktop presentation.
15. **Generative activity principle**: people learn better when they are guided in carrying out generative learning activities during learning.

(Mayer, 2021, p.399–400)

Mayer links these principles to many studies, and experimental studies estimate the effect sizes of each of these principles to be mostly moderate to large. While this certainly qualifies the CTML as a theory that is (a) a cognitive learning theory and (b) grounded in empirical evidence supporting its efficacy, we observe that the applied evidence we have systematically reviewed here has been relatively more equivocal about the effectiveness of multimedia learning in practice. What can be concluded when the ecologically-valid, applied evidence is either too limited or mixed? The effectiveness of multimedia learning is an important question for this review, and is one that we return to in the discussion of the review implications. For now, we note that the extent one deems the above principles to be evidence-based depends on the standard and nature of evidence that is sought. We view the above principles as a strong theoretical starting point for teachers and researchers working in this area but, equally, we hold that testing and applying such principles in ecologically valid studies and within teacher's practice is likely to reveal both expected and unexpected practical and pedagogical challenges in testing or realising the benefits of multimedia learning. In the implementation section immediately below, we explore some of the practical successes and challenges teachers have reported in our interviews and questionnaires.

Before this, we briefly summarise several studies in our wider evidence grouped under the general heading 'multimedia learning', where this is contrasted with 'traditional' teaching. There are many principles discussed that are apparent across these studies. Some of the studies explored whether the principles of the multimedia theory are also current in hypermedia environments, where the learners have more control and participation:

- **Gerjets et al. (2009)** investigated how learner control affects performance in hypermedia environments when learning about probability theory, and how learners' prior knowledge moderates its possible impact. They found that the high level of learner control positively contributed to instructional effectiveness regarding intuitive knowledge, but also increased learning time and thereby made the instruction less efficient. Based on this, they argue that 'the idea to use multimedia design principles for hypermedia learning is too simple and that the

benefits and drawbacks of learner control depend heavily on learning objectives and time constraints' (p.360).

Some examined generative learning principles in multimedia environments:

- **Killi (2006)** also looked at the impact of a participatory multimedia learning model where learners produced part of the learning materials themselves. The model aimed to represent the human information processing system and support the transformation of free cognitive resources into germane cognitive load needed for knowledge construction. The paper also elaborated on the results of an empirical study examining the effectiveness of student-generated illustrations. Finnish elementary school students (N = 187) learned about the human immune system by interacting with multimedia learning materials: students performed better on a retention test when they generated their own illustrations by drawing, and when explanations were presented as animations, compared to students who received only textual material or generated illustrations from images offered.
- Also exploring the element of interactivity and participation, **Moreno et al. (2001**) investigated the learning of college students (in Experiment 1) and seventh-grade students (in Experiment 2) through a computer-based multimedia lesson with various degrees of interaction. They found that students who themselves participated remembered more and transferred their learning more compared to those who had not participated. They also conducted two experiments where the 'pedagogical agent' was either presenting them with material as speech, on-screen text, with an image of the agent on the screen, or in a video of a human face. They found that students who had received material via the speech, image, and human face agent had better retention and problem-solving transfer when words were presented as speech rather than on-screen text. However, the visual presentation of the agent did not affect test performance. Therefore, they argue, interactive pedagogical agents can usefully be introduced to communicate with students via speech to promote meaningful learning in multimedia lessons.

One examined several CTML principles:

- **Kutbay and Akpinar (2021)** studied the effects of modality, redundancy, and signalling principles, considering abstract and concrete representations of an animation of electricity unit in real middle school settings. Their study recommends that when developing material for middle school students, narration is preferable to on-screen text as it reduces extraneous cognitive load. They also found that redundancy and signalling did not have either a significantly positive or negative effect on learning, perhaps because the students who studied with redundant instruction ignored the written text representation and hence avoided cognitive overload. With regards to concrete versus abstract representations, they argue that 'children do not always need concrete representation to understand science concepts, and […] sometimes, using an abstract representation may be beneficial to students' learning [...] multimedia designers and teachers should be aware of the degree of cognitive load correlated with abstract or concrete representation may change according to learning unit, learning objectives, and students' background knowledge' (p.143).

Implementation

What practical activities, principles, and strategies are teachers using to present multimedia information? What are the practical challenges for teachers presenting multimedia information?

In our discussion of visual aids, we outlined many of the approaches teachers reported using to implement dual coding and related ideas. These were quite varied, based on apparently different aims and methods. In these, and in the teacher interview and questionnaire data more generally, there were many mentions of the use of images, diagrams, icons, visual prompts, drawings, and illustrations. Various subject areas were mentioned, albeit with a higher concentration in maths and science, and to a lesser extent English or vocabulary. This greater concentration of maths and science applications was also evident in our main evidence review. In our systematic review of visual representation and illustration, a range of subjects were mentioned but with a greater concentration in maths and science. For the use of diagrams and spatial learning, the focus was almost exclusively on science and maths (but see the embodied learning area review, next, which arguably relates and has a greater range). Many teachers described dual coding as something they felt was a general and long-standing part of common teacher practice and something applicable to most subject areas.

One prominent dual coding discussion point emerging from our teacher responses was that the strategy was particularly beneficial for low-attaining students, students with special educational needs (SEN), and students with English as an additional language (EAL). A selection of related points (mostly but not entirely supporting this point) is provided below:

- ‘Dual coding including nonverbal cues especially helps my younger EAL students with simple concepts. We have to study complex terms, for example, “omnipotent”, and convey these as simply as possible.’
- ‘Dual coding is excellent for new vocabulary for weaker groups.’
- We use dual coding all the time as we have many EAL pupils and SEN pupils. For example, on Smart Slide we will always include an image and often videos as well.’
- ‘Dual coding also helps all in the class embed new concepts and this is especially beneficial for [students with] SEN.’
- ‘All students benefit from dual coding during explanations, but especially those who are sometimes less engaged. It provides a constant focus and reference point.’
- ‘Spaced practice and dual coding seem to really help those with learning problems.’
- ‘[Dual coding] definitely supports ELL and SEN children but [we] have noticed [an] increase in performance across all subjects, particularly English.’
- ‘Very helpful for students with SEND, for example, explicit instructions very clearly delivered reduces cognitive load.’
- ‘Dual coding works really well in all Key Stages, especially with students who have reading difficulties. It has been less successful with EAL students.’
- ‘I have a SEND child who was working at Year 1 level in maths, reading, and writing. She has made incredible progress and her confidence has soared [since implementing dual coding].’
- ‘I've seen that it can work well for SEN pupils ... at least those with moderate SEN difficulties.’
- ‘SEND pupils benefit from dual coding as well as EAL as it helps develop their English too.’
- ‘Repetition and visual strategies [are] better for SEN learners to process and remember.’
- ‘Pupils with ASD have responded particularly well to dual coding and pre-teaching of complex vocabulary prior to reading a difficult class text.’

- ‘These strategies are great for SEN but they are also great for challenging pupils, which is often not considered.’
- ‘I find dual coding and the ideas involving the reduction of cognitive load particularly useful with classes made up of low-prior-attaining students … I think that it can hold very able pupils back, particularly if they already know a lot about the topic ... Very low ability seem to need even more breaking down of work and simpler, more regularly repeated, use of the same images combined with words.’

Challenges

Teachers also discussed challenges they faced when implementing dual coding. Some of these were related to difficulties understanding the concepts and principles. Some were more practical in nature.

- ‘Dual coding … drawing is my weak point (except for sheep) and I find it difficult to create ideas that aren't just maths graphs.’
- ‘Dual coding … unsure how to implement it in the classroom and I remain unconvinced yet of the value of it versus the amount of time it would take to implement properly. I would welcome CPD on this ... My pupils dislike the pictures I've tried to use—too babyish.’
- ‘Dual coding ... multimedia learning: I don't fully understand the concept so I am not sure if I am implementing it or not.’
- ‘Dual coding … I need more practice at finding suitable non-verbal information to support the verbal or written information.’
- ‘Dual coding seems to be widely misunderstood. It is often confused with 'adding diagrams'. This makes dual coding less widely used as it is better used on the spot to visualise abstract ideas rather than shared widely through picture-based resources.’
- ‘Students are less inclined to think and access information in a different way, for example, pictures. They like the teacher doing it but less inclined to do it for themselves.’
- ‘Just don’t really get the point of dual coding. Seems everyone just slaps a little icon and apparently they learn.’
- ‘Dual coding … I find students struggle to work out how to do this effectively and just end up drawing pictures for the sake of it and it loses the benefit.’
- ‘Dual coding has been bastardised somewhat, reduced to pretty icons and complex graphic organisers.’
- ‘I find it hard to know how to implement and find it too time-consuming.’
- ‘Dual coding is hard as to deliver information without having the writing on the board can be challenging for teachers to remember content. Images/key word whilst teacher [is] talking sounds great in practice if teachers are secure in what they are delivering and also rehearsed, which is a challenge when delivering a wide range of subjects across a day and many of which may not be your speciality.’
- ‘When we are doing dual coding … we might be going through key terms and we might put pictures up to help them. I think that those pictures are helping them to understand the terms of criminology, but quite often I get a student asking, “Why is there a picture of that on the board?” So I think, sometimes, what we are delivering and what they are taking from it are two different things.’ (Interviewee 2)

Final thoughts on this strategy area

Our systematic review of classroom trials found mixed results for visual aids, diagrams (in maths and science), and illustrations and positive results for spatial approaches (in maths). Our reflection on the overall area was that many principles appeared to be at play, with practical challenges to navigate these. The variation in practice, the challenges of the implementation, and the many principles we have discussed in this section support this overall impression. This complexity notwithstanding, there was indicatively supporting evidence in both the main review and the wider evidence reviewed here that dual coding and principles from the cognitive theory of multimedia learning can be employed to good effect.

B7. Embodied learning and physical factors

Overview of area

Introduction to this section

In the scoping and protocol development for this review, we searched for concepts and strategies from cognitive psychology and neuroscience that may have implications for classroom practice. This identified a wide range of strategies informed by cognitive science beyond those more commonly represented in policy and practice sources we reviewed (see scoping and protocol development appendix for a concept map). We did not conduct dedicated searches for the strategies reviewed in these wider areas but nonetheless, several studies were located via more general search terms for learning, memory, and cognitive science.[26] We located a group of studies that met our priority criteria within the more general searches yet fell outside of our focus strategy areas. We made a judgement about the weight of evidence (and whether this was sufficient to reach a judgement on effectiveness) and about the value to the breadth of the review of including studies in these additional areas. These wider study areas fell into two groups:

- **'embodied' learning**, where studies examined questions such as how enacting concepts, gesture, tracing, and actions could support learning; and
- **physical factors** such as exercise, nutrition, and sleep.

For the first area, we decided that its inclusion strengthened the review and that the weight of evidence was sufficient for analysis. We note, however, that the lack of targeted searches means that we are less confident that we have located the majority or that we have an unbiased subset of the relevant literature. We decided that the second group, physical factors, required a dedicated review and briefly described and signpost readers to the studies we located in the discussion and questions section only. Below we expand on these decisions and define each area.

Definitions and review inclusion decision

Physical factors

There were areas of literature (both scientific and professional) that linked physical factors such as exercise, nutrition, and sleep to successful learning. Our scoping suggested that this area of research is likely to have many implications for education; however, we decided against a systematic review. This decision was taken due to (a) the limited evidence in the area in relation to nutrition, (b) the lack of targeted searches combined with our view that there are likely to be many studies in this area not located by our search terms, and because (c) many studies of physical influences on learning had policy rather than classroom implications. We concluded that separate, dedicated studies into specific questions would be more appropriate—for example, studies of the importance of sleep and its implications for school start times or studies on the impact of nutrition on learning and how this can be improved within and outside the school. While we do not review factors relating to physical activity levels, sleep, or nutrition here, there are potentially significant studies to be reviewed, although these

[26] Our searches included the general cognitive science terms: 'cognitive', 'brain', 'neuro', and 'learning science' and general memory terms such as 'working' and 'short-term' memory (related to dual coding and cognitive load, for example). See Appendix 3 for full details of the literature searches.

were outside of our scope. However, we do briefly discuss and signpost readers to several studies that did arise in our discussion and questions section.

Embodied learning

We also located, through general searches, a range of studies relating to 'embodied' learning (or 'embodied cognition'), which examined questions such as how enacting concepts, gesture, tracing, and actions could support learning. Many of these arose from (a) general search terms relating to cognitive science, (b) search terms relating to dual coding, and (c) search terms relating to memory and cognitive load. Embodied learning was also a concept evidence in sources from our scoping review in which authors linked mental and physical processes, for example:

> *'If mental processes can influence physical ones, then the question arises as to whether this happens in reverse. Do our physical selves influence how we think? Explorations in this field come under the banner of embodied cognition. George Lakoff (2015), a major figure in this field, has led the way in revealing how much our thoughts, as represented by language, are bound up in physical metaphors. For example, we might describe our mood as up or down, reflecting how we might feel physically present ourselves as upright when feeling positive and in a more downcast shape when feeling low or 'down' […] Ionescu and Vasc (2014) suggest that embodied cognition implied that concrete experience is also needed to develop a deep grasp of abstract concepts and high-order thinking.'*
>
> *(Tibke, 2019, p.67)*

There were strong connections between this area and studies of multimedia learning and dual (or triple) coding, particularly those with a strong spatial element. The studies that we review below focus on how embodying learning may support conceptual and factual learning. Given sources such as Tibke, we judged this area to be sufficiently distinct and sufficiently important for applied cognitive science to include embodied learning as a separate section and review the evidence we located in this area. This has, however, felt a very exploratory area for the review. Without dedicated searches, we are less confident that our evidence represents all or even most of the studies in this area. We note, however, that this is a rapidly developing and potentially fruitful area of thought and research within the basic cognitive science (Collins, 2019, p.63) and—notwithstanding the limitations of our evidence-base—decided that it is of value to review the applied evidence we have relating to embodied learning strategies.

Overview of the evidence-base

Table B7.1: Embodied learning and physical factors—overview of study priority ratings

Priority Level	Overall rating	Ecological validity	Relevance and definition for focus CS practices	Added value to evidence-base
High	1	1	2	1
Medium	13	13	20	13
Low	12	12	4	12

The review study database contained 26 studies in the embodied learning category. Of these, 14 were graded as being of sufficient ecological validity, relevance, and value for inclusion within this analysis of the evidence (high and medium). One study scored highly across these criteria and was identified as *potentially* providing strong evidence in this area (high).

As discussed above, we did not specifically search for embodied learning studies; reflecting this, many studies were graded as medium relevance, usually loosely fitting our definitions around multimedia and dual coded learning. However, with this additional area of analysis, many might be considered a strong fit for embodied learning. We have retained our pre-planned definitions for purposes of transparency. In this area, we judged only one piece to have high ecological validity as there were many contrived, researcher-led, and highly scripted interventions. We discuss this further below.

In this area, we have identified one general group with sufficient evidence to examine the effectiveness of the strategy:

- **embodied cognition**—14 studies, of which one is graded as high priority and thereby identified for in-depth analysis.

As this section was an offshoot of the dual coding and other related areas, all 14 medium and high papers in this area were included in the general strategy group; there were no studies in wider areas, although, as we note above, we discuss how physical factors can affect cognition in the discussion and question section.

Main findings

Strategy 13: Embodied learning and physical approaches

Concise definition

Embodied learning and physical approaches involves enacting or representing concepts through movement or the body, including learning or enhancing learning through the use of the body's sensory or motor capabilities.

Full definition and description

Embodied learning and physical approaches involves enacting or representing concepts through movement or the body, including learning or enhancing learning through the use of the body's sensory or motor capabilities. In this group we included studies relating to physically doing, experiencing, or acting out or playing with the learning object, including both concrete and representational approaches, and in particular gesture and actions.

Selected examples

Examples of this strategy from our database include:

- Teachers used specific hand gestures to refer to two sides of an equation in Cook, Duffy and Fenn (2013). 'Whenever she said the words "one side," she swept her left hand back and forth beneath the left half of the equation and when referring to "the other side," she swept her right hand back and forth beneath the right half of the equation' (p.1866).
- Margolin et al. (2020) used an app which visualised key physics curriculum concepts relating to motion, force, and energy based on students' playground movements and play (for example, plotting in an information dashboard the distance travelled, speed, and time for a child doing cartwheels across the playground).

- Hu, Ginns and Bobis (2014) and Ginns et al. (2015) asked learners to trace out elements of geometry worked examples with their index finger. Also, Tang, Ginns and Jacobson (2019) asked children to trace with their index finger key flows within the water cycle diagram.
- Young children represented number through movement in Ruiter, Loyens and Pass (2015), for example, demonstrating counting in steps (10, 20, 30, 35, 36) using three large jumps, one medium, and one small.
- Mavilidi et al. (2015) used gestures to represent words to support foreign language vocabulary-learning.
- Students were taught gestures to understand 'opposing forces' when learning about tectonic plates in Kaschak, Connor and Dombek (2016).

Evidence for this approach

There were 14 studies for embodied learning and physical approaches. Of these, one was graded as high relevance and quality. Full details of all medium and high studies are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed for this strategy are characterised as follows:

- **Pupil age and characteristics**. The age of students in this area ranged from the early years (age five) to middle-school age children (age 14). There was a good spread of studies within these age ranges, particularly between 5 and 12 years old. These results therefore mostly represent primary age children, with some potential relevance for early secondary children.
- **Location**. Many studies in the area were conducted in the U.S. (four) or Australia (four). Other countries represented in the evidence were Cyprus (two), Iran (one), Taiwan (one), the Netherlands (one), and Switzerland (one).
- **Learning areas**. In this area, there were four studies of maths looking at topics including number knowledge, geometry, and general maths test scores. There were eight studies relating to language, including reading comprehension and vocabulary; one of these was foreign language learning. Three studies related to science, examining physics, plant knowledge, and the water cycle.
- **Outcome measures**: Ten out of the 14 studies used a test designed by researchers aligned to the targeted learning content. In addition, three used standardised instruments and one used a researcher-designed test using items from state assessments.
- **Design and delivery**. The majority of interventions were designed and delivered by researchers (10 of 14). There was one study delivered by teachers, but not the regular class teacher, and one delivered by researchers aided by teachers. Two studies used educational (electronic) games that teachers were trained on to help students facilitate completion. Ecological validity in this area was, therefore, relatively poor. We also note that many of the interventions were relatively short in duration with six being delivered in a single experimental session of 50 minutes or less.

High priority studies in this area

There was one study in the embodied learning and physical approaches category that was rated as having high strength and validity of evidence. We conducted in-depth analysis of this study and have completed a full risk of bias assessment, summarised in the appendix.

Margolin et al. (2020). This study investigated the effects of a play-based middle school physics programme on physics knowledge. This was a randomised experiment with conditions randomised at class level. There was significant attrition: from over 3,000 sixth-, seventh-, and eighth-grade students

and 60 teachers from 50 schools originally recruited to the study, results were obtained for 1,197 children and 45 teachers. The intervention tests the Playground Physics programme—an app with associated curriculum resources designed to allow students to connect complex physics concepts to what they do in real life. The intervention was provided in addition to regular curriculum, which was the control condition. The intervention was a six-week supplemental physical science curriculum comprising three units, energy, force, and motion (each unit six to seven hours over one to two weeks) with content developed by researchers but aligned to the state curriculum. The outcome measures were pre- and post-tests of selected items from publicly available state assessment items as well as research-based instruments.

Key findings. In terms of the results, students who were taught the Playground Physics curriculum had a higher score on the post-test assessment of physics knowledge than those taught the regular physics curriculum (g = 0.38). Our risk of bias assessment flagged 'some concerns' with the high attrition reported above and potential selection of reported results. The overall risk of bias judgement was 'some concerns'.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section we focus on the study outcomes, summarised in Table B7.2. The study identified as high relevance and quality is marked with an asterisk.

Table B7.2: Embodied learning and physical approaches—summary of evidence

Study	Focus	Population	Findings
High Priority Studies			
*Margolin *et al.* (2020)	Effects of a play-based middle school physics program on physics knowledge	*N* = 1,197 Grades 6,7,8 50 schools, 45+classes US	***Positive*** • Students taught Playground Physics curriculum had higher score on the post-test assessment of physics knowledge than those taught regular physics curriculum (d = 0.37, 95 % CI = 0.26, 0.50).
Larger Studies (pupil n > 500) (Medium Priority)			
There were no larger studies at the medium priority level			
Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
Badinlou *et al.* (2018)	Effect of enactment cues on recall of action phrases	*N* = 410 Ages 8-14 Unknown schools/classes Iran? (schools affiliated with Education Organization of Tehran)	***Positive*** • Support for 'enactment effect': Enacted encoding had a recall advantage over verbal encoding regardless of the cue manipulations, • Presenting objects and semantic-integrated items can moderate the enactment effect
Cook *et al.* (2013)	Effect of gestures on mathematical equivalence knowledge	*N* = 184 Grades 2-4 7 schools, 22 classes US	***Positive*** • Main effect of gesture condition on each test, with children in the speech and gesture group performing better than the speech-alone group on immediate Post-test (b=3.12,z=3.24,p<.01), delayed Post-test (b=3.13,z=3.97,p<.0001), and the transfer test (b=2.24,z=2.66,p<.01)
Corcoran *et al.* (2018)	Effect of embodied cognition (Mark DeGarmo Dance program) on reading achievement	*N* = 169 Grade 4 4 elementary schools, 13 classes US	*(no control group) reported for info on feasibility* • Statistically significant difference in reading scores from pre-test (*M* = 284.53, *SD* = 30.82) to post-test (*M* = 295.57, *SD* = 27.45) • The study did not have a control group (single-case design).

Ginns *et al*. (2016)	Effect of tracing worked examples on maths test scores	Australia *N* = 52 / 54 Expt. 1 11-13 years 2 schools Expt. 2 *M* age = 9.3 years 1 school	***Positive*** • Students in the tracing condition outperformed the non-tracing condition on transfer problems in both Experiment 1 (d = .78, 95 % CI = 0.21, 1.34) and Experiment 2 (d = .50, 95 % CI = 0.02, 0.98) • Hypotheses regarding self-reports of cognitive load were not supported
Hsaio *et al*. (2018)	Effect of gesture on plant knowledge and motor skills	*N* = 142 5-6 years 1 kindergarten, 8 classes Taiwan	***Positive*** • Gesture group achieved significantly better scores than the traditional learning group (*d* = 0.35, 95 % CI = 0.02, 0.68). • Gesture group also achieved higher scores on motor skills
Kosmas *et al*. (2020)	Effect of embodied learning on expressive vocabulary	*N* = 118 7-10 years 6 primary schools Cyprus	*(no control group) reported for info on feasibility* • Expressive vocabulary scores increased significantly from pre to post tests (d = 0.65)
Mavilidi *et al*. (2015)	Effects of whole-body movements (exercise) and part-body movements (gesture) on foreign language vocabulary performance	*N* = 125 *M* age = 4.94 years 15 childcare centres Australia	***Positive*** • Children showed highest scores in the task-relevant physical exercise group (where they used physical exercises to enact words to be learned) - they outperformed all other conditions for free-recall performance (0.82, 95 % CI = 0.28, 1.37) • Similar results obtained for cued recall, but no difference between both whole-body physical exercise conditions
Ruiter *et al*. (2015)	Effect of body movement on number knowledge	*N* = 118 *M* age = 7.10 years 2 elementary schools The Netherlands	***Positive*** • When comparing the movement conditions (M=8.23, SD=2.21) with the control conditions (M=7.18, SD=2.65), having math training with movements significantly increased performance compared to the non-movement (d = 0.43, 95 % CI = 0.06,0.80). • Therefore, embodied learning improved number knowledge, but additionally including self-observation made no difference
Schmidt *et al*. (2019)	Effect of embodied learning on foreign language vocabulary learning	*N* = 104 *M* age = 9.04 years 6 elementary school classes Switzerland	***Positive*** • Both the embodied learning (d = 1.12) and the physical activity condition (d = 0.51) were more effective in teaching children new words than the control condition • No difference between embodied learning and physical activity groups on memory performance
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Hu *et al*. (2014)	Effect of tracing worked examples on geometry learning	*N* = 56 11-12 years 2 schools Australia	***Positive*** • Students who traced made fewer errors than those who did not trace (d = 0.54), but a ceiling effect meant this could not be fully analysed (78.6 % students correctly solved all questions)
Kaschak *et al*. (2017)	Effect of gesture ('enacted reading') on abstract text comprehension	*N* = 65 11-12 years 1 school US	***Positive*** • Enacted reading resulted in improved content knowledge for both Grade 3 and Grade 4, for most topic units • These pre-test to post-test changes were generally positive and medium-to-large in magnitude
Kosmas *et al*. (2019)	Effect of embodied learning on expressive vocabulary	*N* = 52 7-10 years 2 primary schools, 4 classes Cyprus	*(no control group) reported for info on feasibility* • Expressive vocabulary scores increased significantly from pre to post-tests (d = 0.28)

Tang *et al*. (2019)	Effect of tracing on knowledge of the water cycle	*N* = 46 9-12 years 1 school Australia	***Positive*** • Post-test scores higher for tracing group than non-tracing group, for both similar (d = 0.74, 95 % CI = 0.13, 1.35), and transfer items (d = 1.11, 95 % CI = 0.42, 1.70). • Self-reported extraneous cognitive load rating lower for tracing group than the non-tracing group (d = 0.84)

* High priority study identified for in-depth analysis.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in Table B7.3.

Table B7.3: Embodied learning and physical approaches—quality of evidence assessment (based on the GRADE approach)

Strategy	Embodied learning
Number of studies	There are 14 studies in this area of which 11 report causal evidence. Of these, one was rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	Eleven studies are randomised experiments (there are three pre-post-only experiments).
Risk of bias	Our risk of bias assessments on the high-quality papers identified some concerns with attrition. As a result, we judged the study to have 'some concerns' for risk of bias and cannot be entirely confident that results in this area are supported by strong studies with a low risk of bias.
Inconsistency	**Result consistency**. The results were consistently positive. Effective sizes ranged from small to large, making an effect size estimate uncertain, but the probably of a negative overall effect low.
Indirectness	**Practice heterogeneity**. In this study, the main groups of practices relate to (a) gesture, (b) tracing, and (c) physical activity and play. We cannot be confident that these form a suitable homogenous group, especially when comparing a and b to c. There is a potential distinction between gesture as signs and observing and learning through movement. **Population, measure, and outcome heterogeneity**. Our sample spanned the primary age range and into the early secondary range (age 5 to 14). While this restricts the results to these students, this range was well represented. A range of subject areas was represented so supports generalisation of the results as a potentially effective principle of learning while increasing the practice and outcome heterogeneity. The outcome measures were mostly (10 of 14) researcher designed tests aligned to the content. There were some examples of standardised assessment from other studies. **Design and delivery**. The majority of interventions were designed and delivered by researchers (10 of 14). Ecological validity in this area was, therefore, relatively poor. We also note that many (at least six) of the interventions were relatively short.
Imprecision	**Group sizes**. There were several (five) relatively small studies (N < 100), seven small to medium (101–200), and two larger studies (N > 201), with one with over 1000 pupil participants. Higher-precision estimates in this group include: - *Margolin et al. (2020): d = 0.37 (95% CI: 0.26, 0.50); - Hsaio et al. (2018): d = 0.35 (95% CI: 0.02, 0.68); - Mavilidi et al. (2015): d = 0.82 (95% CI: 0.28, 1.37); and - Ruiter et al. (2015): d = 0.43 (95% CI: 0.06, 0.80).
Publication bias	There is a slight suggestion of larger effects for smaller studies (without corresponding small or negative effects on the other side of the distribution).
Other considerations	**Searches**. As we discussed at the start of this section, we have not conducted targeted searches for 'embodied' learning. We cannot be confident that we have a sufficiently large or representative sample of the literature in this area.
Overall confidence	**Low (++)** Our confidence in the effect estimate is limited: the true effect may be substantially different from our estimate.

Confidence reasons	Key reasons for downgrading the certainty for evidence include: - limited quantity of evidence in this area, especially when factoring in that this group includes several related approaches (gesture, tracing, and physical activity and play); - there was only one high priority study in this area; some issues were identified in the risk of bias analysis; and - the majority of interventions were designed and delivered by researchers (10 of 14); ecological validity in this area was, therefore, relatively poor.

Summary of findings for this strategy

Main finding. Evidence in this area is consistently positive with a range of small to large effects estimated. The evidence was quite limited but suggests promise for gesture, tracing, and physical activity and play.

Estimated impact. Higher precision estimates range from d = 0.35 to 0.82. The study that we judge to have the highest precision estimate, Margolin et al. (2020), despite some concerns raised relating to attrition in the risk of bias analysis, estimated an effect of d = 0.37 (95% CI: 0.26, 0.50).

Confidence in impact estimate. Our confidence in the findings and effect estimate in this area is low due to limited quantity of evidence, especially when factoring in that this group includes several related approaches (gesture, tracing, and physical activity and play). Moreover, the majority of interventions were designed and delivered by researchers (10 of 14). Ecological validity in this area was, therefore, relatively poor.

Heterogeneity. This group includes several related approaches (gesture, tracing, and physical activity and play). Studies covered a range of subject areas but were all of primary age or early secondary age children, limiting this finding to this specific age range. We have not conducted further analysis of effect differences related to these factors due to insufficient evidence.

Other points. There were three pre-experiments; although these did not have a comparison group, all reported increased learning from pre- to post-tests.

Embodied learning—overall evidence summary and conclusions

Summary of results

In this section, we have reviewed 14 studies focused on embodied learning, which we grouped into a general strategy group for review. Our results for these are summarised in Table B7.4.

Table B7.4: Embodied learning—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Embodied learning	Fourteen, of which one was graded as high priority.[1]	Evidence in this area is consistently positive with a range of small to large effects estimated. The evidence was quite limited but suggests promise for gesture, tracing, and physical activity and play.	Our sample spanned the primary age range and into the early secondary range (age 5 to 14). A range of subject areas were represented providing a tentative suggestion of more general applicability across subjects.	Low (++)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Conclusions about strategies in this area

Embodied learning

Our headline conclusions in this area are:

- Eleven studies in this area reported causal evidence, all of which found embodied learning more effective than a control group.
- Embodied approaches to learning show promise for primary and early-secondary education.
- Many studies found moderate to high effect sizes as well as some smaller positive results. The single study rated as high relevance and quality in this area was Margolin et al. (2020) who found an effect size of *d* = 0.37 in a study of embodied and play-based learning in physics compared to the normal physics curriculum.
- The evidence-based in this area was, however, limited. In particular, there were specific issues with ecological validity for studies in this group with most interventions being researcher-designed and delivered.
- This, and the potential limitations stemming from lack of targeted searches in this area (studies were located through more general search terms), lead us to rate our confidence in these conclusions as low.

Evidence-informed discussion and questions

About this section

As discussed at the start of this section, embodied learning was an area in which we did not conduct targeted searches. However, we have reported all studies of embodied learning in the main section. This section, therefore, presents a short discussion of these along with teacher perspectives. We also briefly summarise some of the indicative evidence for an influence of physical factors on cognition.

Embodied and spatial cognition

Embodied cognition is a very recent development in cognitive psychology. We have not targeted searches for this within either the main review or the practice review. Instead, we refer readers to accounts of embodied cognition such as Shapiro (2019) and Wilson (2002) for more information. Here, we briefly report related ideas from the practice review linked to embodied cognition. In the evidence review above, we note that we have loosely brought together studies relating to embodiment and physical aspects to learning. We noted at the start of this section that many of these studies arose in searches due to links with multimedia learning and dual (or triple) coding, particularly those with a strong spatial element. The other link was with cognitive load. In their practitioner-focused account of generative learning (see Working with Schemas section), Enser and Enser (2020) discuss learning by enactment. They claim, citing Paas and Sweller (2012), that enactment, of which gesture is one example, links to cognitive load theory and—albeit with weaker evidential support—fits with the overall theory of generative learning. One particular area where embodied learning's potential might be explored is for younger children who might benefit from the greater concreteness that comes with gesture and enactment in learning.

In interviews and questionnaires, there were a small number of examples related to this area:

- 'I mostly teach languages in school and I use a lot of actions ... I have done some action research on a very simplistic and small scale as I was interested in whether they really helped children learn words better. They do appear to. I have use associative actions, for eample, remind them of the word, like hands by ears for listen (*écoutez*) or whiskers for a cat ... I think clear movements seem to work best but I haven't actually tested it. Just years of experience! We came to the conclusion

[that] it is a type of dual coding ... I have often wondered if the adoption of a more kinaesthetic approach to language learning might make the vocabulary introduced more liable to stick but have never got further than an aerobic session to introduce and practice prepositions in German!'

- 'In Early Years, play-based pedagogies are aligned to the development of executive functions, the study of which is also cognitive science. There is a lot of evidence connecting play and the development of executive functions, but play is often seen as at odds with 'scientific' ways of teaching.'
- '[I'm] really interested in embodied cognition and saw that it was mentioned recently by John Sweller in a paper reviewing cog load theory ... but don't know a lot about it.'
- 'Embodied cognition [is] especially [useful] for teaching vocabulary.'
- 'Dual coding seems to help all but some appear more interested than others. It still helps older ones too, especially actions, but some don't think it [is] "cool" (they get over it usually and allow themselves to join in!).'

Embodied learning therefore seems to connect to many of the focus cognitive science strategies as a form of encoding or representing information and a factor for cognitive load. In their recent review, Sweller and colleagues summarises this as follows:

> *Research supporting the embodied cognition view shows that observing or making gestures leads to richer encoding and therefore richer cognitive representations. Interestingly, the involvement of the more basic motor system seems to reduce load on working memory during instruction (for example, Goldin-Meadow et al. 2001), which means that this richer encoding is less cognitively demanding and which confirms the evolutionary account of cognitive load theory.*
>
> *(Sweller* et al., *2019, p.286)*

Physical factors

In this area's opening section, we described how the review had located areas of literature (both scientific and professional) that linked physical factors such as exercise, nutrition, and sleep to cognition and ideas in cognitive science. Our scoping suggested that this area of research is likely to have many implications for education; however, we decided against a systematic review. This decision was taken due to (a) the limited evidence in the area, in particular in relation to nutrition, (b) the lack of targeted searches combined with our view that there are likely to be many studies in this area not located by our search terms, and because (c) many studies of physical influences on learning had policy rather than classroom implications and (d) did not frame physical factors within a cognitive science framework. We concluded that separate, dedicated studies into specific questions would be more appropriate, for example, studies into the importance of sleep and its implications for school start times or studies on the impact of nutrition on learning and how this can be improved within and outside the school.

While we do not review factors relating to physical activity levels, sleep, or nutrition here, we do briefly report teacher perspectives in this area followed by signposting readers to several studies that did arise.

In terms of practitioner perspectives, there were several relevant comments. These are from the questionnaire, mostly in response to us asking whether there were another other areas of cognitive science that were important beyond the focus strategy areas.

- ‘Exercise at an early point in the day creates ‘stress’ on the brain; this ‘stress’ means that the brain is comparatively less stressed during a school day. This is a good reason for morning/first thing exercise.’
- ‘RCTs have been done on exercise programmes to improve academic outcomes.’
- ‘Teaching students from Year 7 about the brain and how it learns—the importance of sleep, hydration and concentration.’
- ‘I introduced the “running a mile a day” at my school to increase fitness levels and raise attainment in the classroom.’
- ‘The Daily Mile.’
- ‘Movement breaks.’

We have included references for the 24 studies (including some protocols) we identified in the Embodied Learning Appendix (Appendix 11). In addition, below we provide a brief summary of six studies that, on inspection, we identified as providing particularly strong evidence in this area.

Table B7.5: Physical factors for learning—high-quality studies

Short reference	Focus	Sample	Finding
Have et al. (2018)	The effect on children of integrating physical activity into maths lessons.	Twelve Danish schools. A total of 505 children with mean age 7.2 ± 0.3 years.	Children in the intervention group improved their maths score by 1.2 (95% CI: 0.3, 2.1) more than the control group (p = 0.011). Relative Cohen’s *d* effect size for group differences in the change in maths scores was *d* = 0.38. However, the intervention did not affect executive functions, fitness, or body mass index.
Mullender-Wijnsma et al. (2016)	The effects of an innovative physically active academic intervention on achievement.	499 children (mean age 8.1) from 2nd- and 3rd-grade classes in 12 elementary schools.	After two years, the intervention group had significantly greater gains in a mathematics speed test (P < 0.001; effect size [ES] 0.51), general mathematics (P < 0.001; ES 0.42), and spelling (P < 0.001; ES 0.45). No differences were found on the reading test.
Husain et al. (2019)	Fit to Study aimed to increase the amount of physical activity undertaken by Year 8 children in PE lessons.	Randomised controlled trial involving 104 schools and 8,707 pupils.	There is no evidence that Fit to Study had an impact on Year 8 pupils’ maths outcomes. This result has a low security rating. Attendance at the initial face-to-face training was poor and there were implementation issues that may have affected the results.
Fedewa et al. (2015)	This study explored whether additional curricular physical activity during the school day resulted in gains for children’s fluid intelligence and achievement outcomes.	Participants were children (N = 460) from four urban schools in the Southeast United States.	Results from the one-year study show positive effects for children’s mathematics and reading achievement but no differences across treatment and control groups for children’s fluid intelligence scores.
Bunketorp et al. (2015)	A curriculum-based physical activity intervention on children’s academic	Quasi-experimental design. National test results from 545 students, 122 in the	Curriculum-based physical activity in school may improve the academic achievement and psychological health of children, particularly for girls. Girls attending the

	achievement, wellbeing, health-related quality of life, fitness, and structural development of the brain.	intervention school, and 423 in 3 control schools.	intervention school were more likely to pass national tests in Swedish (odds ratio 5.7) and mathematics (odds ratio 3.2).
Tarp et al. (2016)	The effectiveness of a school-based physical activity intervention in enhancing cognitive performance in 12–14 years old adolescents	Seven intervention and seven control schools. A total of 632 students, mean age 12.9 (SD 0.6).	No significant difference in change, comparing the intervention group to the control group, was observed on the primary outcomes (p's > 0.05) or mathematics skills (p > 0.05).

Final thoughts on this strategy area

Embodied learning is an area that, in our view, needs to be brought more into focus within accounts of cognitive science. Many of the practice-focused cognitive science sources we consulted provided an account of the other strategy areas we have reviewed but were conspicuously silent when it came to embodied and physical aspects to learning, which seem to have received far less emphasis across the practice review literature. Whether this stems from differing influence and concerns of cognitive psychology and neuroscience, respectively, or a more fundamental cartesian dualism (distinguishing body and mind) within popular culture, the separation of embodied and physical aspects of learning from mental aspects of learning does not appear justified by the evidence or what we know about cognition.

B8. Mixed strategy programmes

Overview of area

Definitions

A relatively small group of studies evaluated programmes where two or more of our focus cognitive science strategies were combined. Where studies separated strategies through, for example, multi-arm trials or multiple experiments, we were able to include the studies in our analysis of specific areas. Where only combined results were reported for the effect of multiple cognitive science concepts, we have included this here as a mixed strategy programme.

From the perspective of assessing individual cognitive science approaches, mixed programmes with combined analysis are not ideal. There were, however, several studies in this category that we rated as having high ecological validity (that is, designed or delivered by teachers in classroom settings). Furthermore, programmes of professional development, initial training, or curriculum development are, in practice, highly likely to incorporate a larger range of principles and techniques from cognitive science and beyond. Therefore, studies that evaluate attempts to apply multiple strategies, and especially those that do so at scale, are of great interest.

Overview of the evidence-base

Table B8.1: Mixed strategy programmes—overview of study priority ratings

Priority level	Overall rating	Ecological validity	Relevance and definition for focus CS practices	Added value to evidence-base
High	5	8	6	2
Medium	3	5	2	9
Low	7	2	7	4

The review study database contained 15 studies in the mixed strategy programme category. Of these, eight were graded as being of sufficient ecological validity, relevance, and value for inclusion within this analysis of the evidence (high and medium). Five studies scored highly across these criteria and were identified as *potentially* providing strong evidence in this area (high). There were seven substantive trials reported: two of the studies report, respectively, the effects and implementation of the same programme.

Studies in this category had relatively high ecological validity as many were programmes of professional development or curriculum development specifically designed to apply cognitive science principles in practice (rather than provide proof of the principle or its efficacy). Moreover, several studies involved large numbers of schools, with four large studies, two moderate, and one smaller.

Main findings

Strategy 14: Mixed strategy programmes—general

Concise definition

A mixed strategy programme is an intervention in which two or more cognitive strategies are combined in a single intervention *and* the study was designed in a way that prevented calculation of independent effects.

Full definition and description

A mixed strategy programme is an intervention in which two or more cognitive strategies are combined in a single intervention *and* the study was designed in a way that prevented calculation of independent effects. Programmes of professional development, initial training, or curriculum development are, in practice, highly likely to incorporate a larger range of principles and techniques from cognitive science and beyond. The content of individual programmes in our database are detailed at greater length in the discussion section.

Evidence for this approach

There were eight publications reporting tests of mixed strategy programmes reporting seven substantive studies. Of these, five were graded as high relevance and quality. Note that two studies, Schunn et al. (2018) and Desimone and Hill (2017), report the effectiveness and implementation, respectively, of the same overall programme.

Full details of all medium and high studies are contained in the summary table in the appendix associated with this section.

In overview, the studies reviewed in this area are characterised as follows:

- **Pupil age and characteristics**: All students within these studies were middle school students aged 11 to 14 (Years 7 to 9, Grades 6 to 8).
- **Location**. All studies were either from the U.K. (two) or U.S. (five).
- **Learning areas**. This is another area, as with some of the previous strategies, where the evidence is dominated by a focus on science and maths learning. There were five studies of science and two in maths (one focused on number lines and one on the mathematics curriculum more generally).
- **Outcome measures**. In general, studies combined curriculum-focused questions with standardised tests or measures designed through selecting items from standardised tests (for example, NAEP, TIMSS, or established curricular tests), striking a balance between test sensitivity, curriculum alignment, and rigour. This applies to four out of seven studies. The other three used tests aligned to the focus curriculum.
- **Design and delivery**. Relative to other areas, ecological validity in this area was good. Five of the studies were based on a model of providing professional development or curriculum resources to teachers with regular class teachers then delivering the programme. One study (Feddern et al., 2018) used revision software designed by the researchers with students working independently. In another (Barbieri et al., 2019), the intervention was taught by six researcher–instructors with the carefully scripted lessons to increase fidelity.

High priority studies in this area

There were four studies in the mixed programme category that were rated as high priority. We conducted in-depth analysis of these studies and have completed a full risk of bias assessment, summarised in the appendix.

Cromley et al. (2016) examined the effect of a cognitive science informed curriculum including teaching diagram comprehension in biology. This was an RCT with a teacher-level assignment involving 9,611 seventh-and eighth-grade students of 129 teachers in the U.S. Teachers were randomly assigned to one of three groups: business-as-usual control, content-only, and cognitive-science-based. The cognitive-science-based intervention incorporated three major components (visualization exercises, case comparisons focused on highlighting key science concepts, and spaced testing in the form of daily warm-up quizzes) that were interleaved into the same base unit (Holt Introduction to Matter, Cells or Inside the Restless Earth; FOSS Diversity of Life, Weather and Water, or Earth History). A fourth principle, confronting misconceptions, also informed the design.

All seventh-grade science teachers were assigned to the same condition within each school for two consecutive years provided that they remained employed as science teachers at that same school. Before implementing a modified unit, cognitive-science-based teachers attended three paid days of summer professional development per unit they were implementing. This was coupled with providing supportive material and school year teacher discussion. Teachers in the business-as-usual control condition received neither professional development nor the modified curriculum. Instead, students attended their scheduled classes, completed only the activities included in the standard curriculum, and then completed an end-of-unit test. To measure the outcomes, six sets of three diagram-specific items each were created for each curriculum and added onto the science content knowledge measure to create six unique test forms. These six test forms were then randomly given to students in the study.

Key findings. The cognitive science curriculum group outperformed the content-only and business-as-usual groups. The Cohen's d effect sizes for cognitive science versus content only across six curriculum units were d = 0.48, 0.49, 0.62, 0.52, 0.20, and 0.21. The corresponding effect sizes for the cognitive science informed curriculum compared to the control across six curriculum units were d = 0.52, 0.41, 0.55, 0.11, 0.06, and -0.13. Study 2 examined items with diagrams specifically, with the same pattern of results. The intervention was more successful in classrooms where the teacher was teaching with interventions for the second time, suggesting some practice in implementing the diagrammatic interventions is useful. Our risk of bias analysis identified 'some concerns' with potential selection of reported results due to the lack of a formal pre-planning of analysis. In all other areas the risk of bias was rated as 'low'.

Davenport et al. (2020). This study evaluated an intervention where cognitive science concepts were applied to revise a widely-used middle school mathematics curriculum in the U.S. A between-subjects, cluster-randomised trial design was employed with random assignment conducted at the school level. The study recruited 114 schools and 181 teachers in 22 states were fully enrolled in the study and randomised into an experimental group. There were issues of high attrition with, for example, wider issues of teacher turnover affecting the consistency of the sample. Results are based on 62 schools, 88 teachers, and 2,595 seventh-grade students who participated during the second year. The Connected Mathematics curriculum (CMP2) was revised using the following principles:

- visual and verbal mapping (dual coding);
- worked examples with self explanation;
- space learning; and

- formative assessment (quizzes).

Treatment teachers attended a professional development session at the beginning of each study year and three online follow-up sessions distributed throughout each study year. Treatment teachers taught the CMP2 curriculum using the redesigned student booklets, followed the recommendations for spacing and formative assessment in the teacher guide supplement, and used their practices during the professional development. Control teachers taught the CMP2 curriculum as they had before the study, using the same business-as-usual materials and practices. Outcomes were assessed using a pre-algebra readiness diagnostic developed by the Mathematics Diagnostic Testing Project (MDTP, 2004) administered at the beginning and end of each study year. There were also project-developed unit assessments aligned to the curriculum. These draw on the CMP curriculum and assessment materials and released items from standardized tests (NAEP, TIMSS, MCAS).

Key findings. The results indicated that the treatment group exhibited a higher than expected post-test scaled score than the control group However, the differences between groups were not statistically significant. The Hedges g effect-size estimate for the expected difference on the summative MDTP assessment was 0.12 (95% CI: -0.31, 0.55). Curriculum unit by unit effect sizes ranged from 0.08 to 0.50, although only one result was statistically significant. This review's risk of bias assessment for this study raised concerns about potential deviation from the intended intervention, missing data (relating to attrition), and reporting of results, the latter driven mostly by the specific requirements of the RoB2 analysis than a specific concern. Overall, the study was graded as 'some concern', which we interpret as requiring caution but nonetheless providing indicative evidence.

Yang et al. (2020). This study compared training focused on cognitive science principles applied to the science curriculum against two control conditions: a content knowledge only condition and a business-as-usual condition. The study design was a Cluster RCT in which schools were randomised into one of the three arms: the CS arm, the content arm, and the control arm. For Cohort 1, the final analysis file contained 6,410 seventh- and eighth-grade students in 90 schools with 145 teachers. For Cohort 2, the final analysis file contained 5,508 students in 82 schools with 130 teachers. All schools were from a large urban area in the U.S.

In the study, the 90 schools were randomly assigned into one of three arms: (a) a treatment arm in which the textbook curriculum was modified based on four principles of cognitive science coupled with teacher professional development (PD), (b) a second treatment arm in which teachers received PD designed to improve their knowledge of the science content, and (c) a business-as-usual control group. The researchers provided a 2.5-day (18-hour) summer PD session and four two-hour professional learning community (PLC) sessions during the school year for two years: a total of 34 hours of PD. The PD for both intervention arms happened in the summer and the following school year. Teacher content knowledge was assessed after the summer PD sessions. Ongoing sessions during the school year, which were called 'professional learning community meetings' (PLCs), took place approximately monthly during delivery. The student outcome measures were an end-of-unit test designed by researchers to align with curriculum and the state science test given at the end of eighth grade. The latter had higher stakes and was more likely to represent student's greatest effort but had only moderate content alignment.

Key findings. The analysis-estimated effect sizes for the cognitive science condition compared to the control ranged from 0 to 0.20. The effect sizes for the cognitive science condition compared to the business and usual condition ranged from 0.06 to 0.36. Although some differences were close to being statistically significant, the only significant difference occurred for one unit (of three) for Cohort

2 (of two). For analytical models with total state test scores as the response variable, results were similar to those from the end-of-unit test analyses: the coefficient estimates for cog-sci were all positive, but the differences between cog-sci and control results were not statistically or substantively significant. As discussed, there were various issues with implementation such as might be expected with the implementation of a programme and the associated evaluation at this scale. Our own risk of bias assessment raised some concerns with the randomisation process, deviations from the intended intervention, and selection of reports results (due to the lack of reported information).

Feddern et al. (2019). This study employed a randomised controlled trial to test the effectiveness of biology revision software that incorporates cognitive science principles on biology test scores (spacing, interleaving, retrieval, and visual cues). The trial involved 14-year-old pupils in a U.K. school (n = 829). There were three conditions: first, a 'business as usual' group who completed a 40-minute revision session using a physical guide (massed practice); second, an 'offline' spacing group who completed two 20-minute sessions using a PDF revision guide two weeks apart; third, a software condition using mixed cognitive science strategies and question personalisation based on performance. Students studied independently. The learning measure was a pen-and-paper biology test on the content, consisting of multiple choice, free recall, and short-answer questions.

Key findings. 'Offline' spacing was found to be slightly but not statistically significantly more effective than massed practice; the mixed strategy software condition (M = 8.39) produced significantly higher scores than both. In previous analyses in the spaced practice section, we compared the spaced condition to the massed condition. Here we are concerned with the mixed condition versus the other two conditions, including a single cognitive science principle (spacing) and the other includes none. The risk of bias assessment for this study did not raise any concerns, however, we note that this was published in the Chartered College of Teaching Impact Journal and was at a shorter length, with briefer and less formal reporting, than typical of a journal with a research audience.

Schunn et al. (2018). This study used four principles of cognitive science to make systematic revisions in middle school science instructional modules from two kinds of curriculum: 'textbook science' and 'hands-on science'. Cognitive science principles used in the curricula were:

- identifying misconceptions and student prior knowledge;
- case comparisons;
- visualization exercises; and
- spaced testing.

The study consisted of two randomised controlled trials, one for each curriculum. The textbook curriculum was for students in the seventh- and eighth-grade levels. The study had a sample of 6,400 students in the first year of implementation and 3,200 students in the second, including 229 teachers, 97 schools, in one urban district in the U.S. The 'hands-on' curriculum study included 7,600 students and 4,200 students for the first and second years respectively, with 116 teachers, 65 schools, six urban districts, in two cities in the U.S. Schools were randomly assigned to one of the three arms (cognitive science modifications with professional development, active control with professional development, or business-as-usual). Two cohorts of students were followed in each arm for each setting. In the active control and treatment conditions, teachers received 20 hours of professional development. The outcome measures were end-of-unit assessments, which each involved 18 questions related to the curricular content. The questions for each unit's assessment were developed by sampling items from various item pools (released state tests, released NAEP and TIMSS items; Porter, Polikoff, Barghaus, and Yang, 2013). The implementation was examined in more detail in Desimone and Hill (2017). We

return to discuss this aspect of the programme drawing on both publications in the wider evidence section.

The analysis broke down the results by:

- classes with higher and lower underrepresented minority (URM) student proportions;
- curriculum (textbook based and hands-on);
- unit of study (Cells or Matter); and
- year of study (first and second).

This produced 2 x 2 x 2 x 2 (= 16) overall results.

Key findings

- The results for the **textbook-based curriculum** revealed that the intervention predicted better scores than the control for lower-URM-proportion classrooms (d = 0.21, 0.50, 0.48, 0.52). However, the cognitive intervention did not predict higher scores in higher-URM-proportion classrooms (d = -0.17, -0.09, -0.15, -0.01).
- For the **hands-on curriculum**, in the lower-URM-proportion classrooms, the intervention predicted higher scores in the first year (d = 0.12, 0.22) but not in the second (d = -0.01, 0.04). In higher-URM-proportion classrooms, first year effects were negative or small (d = -0.13, 0.12); in the second year, positive (d = 0.36, 0.18). The control condition (content-only training) had no effect compared to business as usual. In several cases, the content-only training condition scores negatively predicted outcomes (albeit only marginally).

The risk of bias assessment identified concerns with the reporting as the only potential issue. This study had 'low' risk of bias in all other categories.

Overview of all studies in this area

We have reported the overall characteristics of studies for the strategies above. In this section we focus on the study outcomes, summarised in Table B8.2. Studies identified as high relevance and quality have been marked with an asterisk.

Table B8.2: Mixed strategy programmes (general)—summary of evidence

Study	Focus	Population	Finding
High Priority Studies			
Cromley *et al.* (2016)*	Effect of cognitive science informed curriculum including teaching diagram comprehension in biology	N = 9,611 7th and 8th grade students 129 teachers. US	***Positive*** • Cognitive science curriculum group outperformed others. • Cohen's d Effect sizes for cogsci vs. content only across 6 curriculum units: 0.48, 0.49, 0.62, 0.52, 0.20 (ns), 0.21 (ns) • Cohen's d Effect sizes for cogsci vs. control across 6 curriculum units: 0.52, 0.41 (ns), 0.55, 0.11 (ns), 0.06 (ns), -0.13 (ns) (where, ns = not significant at the 0.05 level) • Study 2 examined items with diagrams specifically, with the same pattern of results. • The intervention was more successful in classrooms where the teacher was teaching with our interventions for the second time, suggesting some practice in implementing the diagrammatic interventions is useful.

Davenport *et al*. (2020)*	Intervention to us CS concepts to revise a widely used middle school mathematics curriculum.	7th Grade students, US Results are based on 62 schools, 88 teachers, and 2,595 students who participated during the second year.	***Neutral*** • The results indicated that the treatment group (M = 0.85, SD = 0.84) exhibited a higher expected post-test scale score than the control group's expected scale score (M = 0.74, SD = 1.01); however, the differences between groups were not statistically significant (p = 0.33). g = 0.12 (95 % CI = -0.31, 0.55) was estimated for the expected difference on the summative MDTP assessment (SE = 0.22). Unit by unit effect sizes ranged from 0.08 to 0.50, although only 1 result was statistically significant.
Yang *et al*. (2020)*	A comparison of training focused on cognitive science principles versus content knowledge in science	For Cohort 1: 90 schools, 145 teachers, and 6,410 students. For Cohort 2: 82 schools, 130 teachers, and 5,508 students. 7th and 8th grade students large urban city in the US	***Neutral (near positive)*** • The ESs (d) for Cog-sci versus control ranged from 0 to 0.20. The ESs for Cog-sci versus Content ranged from 0.06 to 0.36. Although some differences were marginally significant, the only significant difference occurred for one unit (of 3) for Cohort 2 (of 2). • For models with total state test scores as the response variable, results were similar to those from the end-of-unit test analyses. The coefficient estimates for Cog-sci were all positive, but the differences between Cog-sci and control were not significant.
Feddern *et al*. (2018)*	Testing the effectiveness of cognitive science-inspired biology revision software (spacing, interleaving, retrieval, visual cues) on biology test scores	• *N* = 829 • Year 9 (13-14 years old • UK	***Positive*** • 'Offline' spacing was no more effective than massed practice , but the software produced significantly higher scores than both controls • Results applied across both selective and non-selective schools. (Effect sizes not provided, nor the information to calculate these).
Schunn *et al*. (2018)* (also see implem-entation evidence in Desimone and Hill (2017))	Four principles of cognitive science were used to make systematic revisions in middle school science instructional modules from two kinds of curriculum	Textbook curriculum N= 6,400, 1st year and 3,200 2nd. 229 teachers, 97 schools. Hands-on curriculum N = 7,600 and 4,200 for the 1st and 2nd years. 116 teachers, 65 schools. 7th and 8th Grade, US	***Neutral / Mixed*** ▪ Textbook-based curriculum intervention predicted better scores than control for lower-URM-proportion classrooms (*d*=.21, p=.152; *d*=.48, p=.097; *d*=.50, p=.012; *d*=.52, p=.010). But not in higher-URM-proportion classrooms (*d*=-0.17,p=.211; *d*=-0.09, p=.43; *d*=-0.15, p=.418; *d*=-0.01, p=.967). ▪ Hands-on curriculum:. In the lower-URM-proportion classrooms, the intervention predicted higher scores in the first year (*d*=.12, p=.251; *d*=.22, p=.03) but not in the second (*d*=-0.01, p=.965; *d*=.04, p=.774). In higher-URM-proportion classrooms 1st year effects were negative or small (*d*=-0.13, p=.288; *d*=.12, p=.251); in the 2nd year positive (*d*=0.36, p=.022; *d*=.18, p=.412) ▪ Control condition (content training). No effect. In several cases, scores were negatively predicted (at least marginally) by this training.
Larger Studies (pupil n > 500) (Medium Priority)			
There were no larger studies at the medium priority level			

Medium-sized Studies (100 < n ≤ 500) (Medium Priority)			
Adey and Shayer (1993)	Concrete activities, cognitive conflict, metacognition, schema development and bridging[27] in science.	N = 424 24 classes Year 7/8, Age 11-13, England	***Positive*** • Some sub-group results negative but mostly positive. 3 of 4 positive in GCSE science for 12+/11+ Boys (ES = 1.03/-0.22) and Girls (ES=0.19/0.67). 3 groups of 4 positive ES in maths, and all in English.
Smaller Studies (pupil n ≤ 100) (Medium Priority)			
Barbieri *et al*. (2019)	Intervention using number lines and 'incorporating key principles from the science of learning'.	N = 51 2 middle schools 6th Grade US	***Positive*** • The experimental group demonstrated significantly more learning than the control group from pre-test to post-test, with meaningful effect sizes on measures of fraction concepts (g = 1.09), number line estimation as measured by percent absolute error (g = -.85), and magnitude comparisons (g = .82). These improvements held at delayed post-test 7 weeks later.

* High priority study identified for in-depth analysis.

Evidence assessment—GRADE analysis

We have appraised the overall evidence in this area using an adaptation of the GRADE evidence appraisal approach. GRADE is not designed specifically for education research. We have reviewed our results against the main evaluation categories, interpreting the guidance for the education context. The results of this assessment are summarised in Table B8.3.

Table B8.3: Mixed strategy programmes (general)—quality of evidence assessment (based on the GRADE approach)

Strategy	Mixed strategy programmes
Number of studies	There are eight publications in this area reporting seven trials. In addition, five of the seven trial effectiveness publications were rated as high priority based on relevance, ecological validity, and added value and underwent in-depth analysis and risk of bias assessment.
Design	All studies are randomised experiments.
Risk of bias	Our risk of bias assessments on the high-quality papers raised some concerns with the randomisation process for one, with deviations from the intended intervention for two (mainly due to implementation issues), and missing data (through attrition) for one. There were gaps in the reporting for four of the five. Four studies were rated as having some concerns overall and one low risk of bias. The latter had limited reporting detail but reported specific aspects required for the risk of bias assessment. We judge, therefore, there to be at least one strong study in this area from the risk of bias analysis. However, our wider judgement on these studies is that this understates the strengths of evidence in this area (with the expectation of tightly controlled experimental conditions for large-scale, real-world evaluations). We hold that all of these studies provide important evidence but that caution is needed given their small number and the risk of bias concerns raised.
Inconsistency	**Result consistency**. The results were mixed, with both positive and neutral results. While some sub-results were negative, the evidence does not suggest that mixed strategy cognitive science programmes are likely to have a negative effect. Positive and neutral results ranged from zero to effect sizes in excess of one (large). Where within this range the expected mean effect for a mixed cognitive science programme is likely to fall cannot be determined by this evidence. We judge the effect to be highly dependent on programme quality and implementation (see below).

[27] i.e., strategies to generalise reasoning to promote transfer.

Indirectness	**Practice heterogeneity**. Studies in this area were mostly focused on improving a curriculum or learning area using CS principles (six of seven studies). The other was similar in that it provided a 'casebook' of written materials and activities along with professional development. Heterogeneity in terms of this general design was therefore high. But, in terms of the underlying organisation of the programmes and the strategies used, there was substantial variation. **Population, measure, and outcome heterogeneity**. The studies were all focused on maths and science in the U.K. or U.S. and students aged 11 to 14. **Design and delivery**. Especially when considered relative to the specific strategy areas we have reviewed, ecological validity is high in this area. Most studies were delivered by teachers and designed to work at scale in real classroom conditions.

Imprecision	**Group sizes**. In terms of pupil numbers, there were four large studies, two medium-sized, and one small; the randomisation for the larger studies was at school level. In general, standardised and appropriate assessments were used. The precision from scale and measurement will therefore be acceptable. There is likely to be high imprecision stemming from treatment fidelity.
	
	Estimates provided by high priority studies were as follows:
	- Cromley et al. (2016):*
	  ○ Cohen's d effect sizes for cogsci vs. content only across six curriculum units: 0.48, 0.49, 0.62, 0.52, 0.20, 0.21 (mean = 0.42),
	  ○ Cohen's d effect sizes for cogsci vs. control across six curriculum units: 0.52, 0.41, 0.55, 0.11, 0.06, -0.13 (mean = 0.25);
	- Davenport et al. (2020):* g = 0.12 (95% CI: -0.31, 0.55); and
	- Yang et al. (2020):* the ESs (d) for Cog-sci versus control ranged from 0 to 0.20. The ESs for Cog-sci versus Content ranged from 0.06 to 0.36. One on comparison was significant.

Publication bias	The medium and small studies in this group were both positive compared to a mixture of positive and neutral results elsewhere.
Other considerations	There were numerous issues with implementation in the group—as discussed further below.
Overall confidence	**Low (++)** Our confidence in the effect estimate is limited: the true effect may be substantially different from our estimate.
Confidence reasons	A low, rather than a moderate, confidence judgement is mostly driven by the small number of studies and questions around their implementation.

Summary of findings for this strategy

Main findings. Overall, the evidence provides a mixture of mixed or neutral results and small to moderate positive results for programmes of mixed cognitive science strategies. There were suggestions in several studies of issues with implementation. We judge the effect to be highly dependent on programme quality and implementation.

Estimated impact. Of the five strongest studies, three had neutral or mixed effects and two, positive effects. Estimates from Cromley et al. (2016) and Yang et al. (2020) had a range between -0.13 and 0.62 and an overall small likely effect size.

Confidence in impact estimate. Our level of confidence was rated as low. A low, rather than a moderate, confidence judgement is mostly driven by the small number of studies and questions around their implementation.

Heterogeneity. One positive result was based on a curriculum redesign delivered at scale through professional development. The other positive result was based on cognitive science principles built into a computer revision programme and, while promising, had lower ecological validity than other studies. There were suggestions in several studies of issues with implementation. We discuss these further below.

Mixed strategy programmes—overall evidence summary and conclusions

Summary of results

Eight publications were reporting tests of mixed strategy programmes; seven of these were substantive studies. Of these, four were graded as high relevance and quality. Note that two studies, Schunn et al. (2018) and Desimone and Hill (2017), report the effectiveness and implementation, respectively, of the same overall programme. Our results for these are summarised in Table B8.4.

Table B8.4: Mixed-strategy programmes—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Mixed strategy programmes	Seven, of which five were graded as high priority.[1]	Overall, the evidence provides a mixture of mixed/neutral to small-moderate positive results for programmes of mixed cognitive science strategies. There were suggestions in several studies of issues with implementation. We judge the effect to be highly dependent on programme quality and implementation.	The studies were all focused on maths and science in the U.K. or U.S. and students aged 11–14.	Low (++)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.
[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Conclusions about strategies in this area

Mixed strategy programmes

Our headline conclusions in this area are:

- Our analysis concerned programmes testing two or more cognitive science principles combined. These programmes typically revolved around curriculum (re)design accompanied by professional development in the cognitive science principles and (to greater and lesser levels of success) implementing the materials.
- Overall, the evidence provides either positive or mixed/neutral results for programmes of mixed cognitive science strategies.
- The evidence presented above shows that, at present, there are few or no large-scale programmes that have been trialled and found to be effective. Moreover, those that have been trialled—of which there are only a small handful of ecologically-valid, rigorous examples—some have yielded disappointing results.
- Of the five strongest studies, three had neutral or mixed effects and two had positive effects. One positive result was based on a curriculum redesign delivered at scale through professional development. The other positive result was based on cognitive science principles built into a computer revision programme and, while promising, had lower ecological validity than other studies.
- There were suggestions in several studies of issues with implementation.
- Our confidence in the effect estimate is low.

Mixed strategy programmes have high potential relevance across the U.K. education system, for all learners and subjects. As (or if) cognitive science strategies are held to be individually effective, combining two or more strategies in a single intervention might be expected to increase the overall impact as multiple, individually-effective strategies combine for collective and additive benefits.

Moreover, as discussed further below, mixed strategy programmes are likely to be an important vehicle for applying and scaling cognitive science informed practices.

Overall, the evidence on mixed strategy programmes presented in this section yields disappointing results. As we discuss at length in the discussion and questions section, our reading of this area is that the effectiveness of these programmes has been determined as much by their programme design and organisation as their underlying teaching and learning principles. Small or null results may stem from the operational issues as much from the (in)effectiveness of the underlying strategies; the evidence we have is not sufficient to support either explanation. There are known issues with implementation in these programmes (as we discuss below). However, these did not apply to all programmes and the relationship between implementation success (in terms of fidelity and dosage) and outcomes is not consistent across the group of studies.

These programmes are likely to draw on principles and practices relating to school improvement, curriculum development, and effective professional development; the success of these at such organisational and policy levels is likely to be important for the success of any intervention at scale. The successful design of school improvement and professional development programmes lies beyond the focus of the present review; nonetheless, we conclude that these will be important considerations if and when cognitive science programmes are tested or delivered at scale.

Evidence-informed discussion and questions

Cognitive science and school improvement

Principles and translation into teacher practice

Here we briefly discuss programme and policy-level features touched on by the larger-scale studies above. It is beyond the purview of this study to discuss these in any detail; we aim to connect some considerations from the evidence in this area to wider literature, policy, and practice.

How does applied cognitive science differ in terms of focus and methods to basic cognitive science? What were the challenges of judging ecological validity?

A good starting point for this discussion is to note that the problem of translating cognitive science principles into teacher practice at scale was not the focus in many studies in previous review sections. Most problematic for inference about transfer and scalability are intervention 'set pieces' delivered by researchers or experts or scripted lessons or computer programmes for independent study. From the perspective of assessing efficacy or experimentally isolating cognitive scientific principles, there is huge value in these studies but for our present purposes of assessing the implications of the evidence for teacher practice, it is important that we also consider the necessary steps to get from a 'proof of principle' to a strategy suitable for widespread implementation by teachers. The ecological validity of studies in the mixed strategy programme area has been relatively high. As a result, below, we discuss some of the factors associated with educational change, drawing on the studies in this area representing some of the most concerted attempts to get cognitive science into practice at scale.

The studies in this section also represent studies across all areas that have conducted an informative test of the science while having a design and emphasis not conducive to advancing our understanding of the issues of implementation discussed further below. Barbieri et al. (2019) is a good example of such a study. This met our broad eligibility criterion of taking place in a classroom in a typical teaching

and learning context. Furthermore, it provides an informative test of a widely applicable question: whether visual representation (in this case, a number line) and cognitive science strategies can improve fraction understanding (see earlier for an outline of the study). Regarding the design and delivery of the actual intervention, the following details are provided:

> - *Each small group was taught by one researcher-instructor. Because of available resources, four researcher–instructors taught one group each and two researcher–instructors taught two groups (one in each school) (p.5).*
> - *The intervention took place during a 6-week period in which all students received specialized help from a teacher within their school. This designated 45-min intervention time is in addition to students' regular mathematics class. In their regular mathematics classes, both schools used the same mathematics curriculum: Connected Mathematics Project (p.8).*
> - *During the additional class period dedicated to intervention, students in the experimental intervention condition received 27 researcher-designed lessons (described further below). These lessons were administered to each of the small groups by one of the trained instructors (p.8).*
> - *Experimental intervention instructors were trained research assistants who also participated in lesson design. Instructors varied in prior teaching experience. Two instructors were doctoral students, two were postdoctoral researchers, and two were previous certified teachers. Each of the six instructors received more than 16 hr of the same training in administration of the lessons from one of the authors of the current paper. Training included practice in use of gestures, proper strategies for providing feedback, instructor/student dialogue, and behavior management. Experimental intervention instructors also practiced teaching the lessons in pairs and provided each other feedback for lesson improvement prior to administering the lessons (p.8).*
>
> *(Barbieri et al., 2019, p.5–8, abridged)*

Within this account, we can identify features that suggest that the intervention might work more widely. These include that half of the delivery team were certified teachers, that the lesson is designed to work in a typical teaching period of 45 minutes (rather than shorter discrete activities), that the intervention was sustained over six weeks and 27 lessons, and that it was built on the regular classroom curriculum, pursuing a typical learning objective for students of this age. However, there are also aspects which create doubt about wider applicability. These include the fact that the lessons were 'set piece' lessons that had been practiced, that the instructors were not (or so we understand from what is reported) the regular class teachers, and that the intervention was delivered in addition to regular mathematics class time. The latter might support inference about this programme being used as a 'catch-up' intervention but renders more difficult conclusions about this approach being preferable in regular teaching to a 'traditional' approach (one not informed by cognitive science) as well as raising questions about feasibility and cost (as separate intervention groups involve extra staffing and compete with other areas of the timetable for the intervention students).

Our aim here is not to reach evaluative judgements about this particular study but rather to surface and illustrate a set of questions present throughout the review when screening for eligibility (and in particular our ecological validity judgement) and within the analysis. There were some studies that provided empirical evidence about whether cognitive science principles might work in authentic

classroom conditions. Ecologically valid studies were designed such that (a) researcher involvement was limited to providing resource and training to support regular teachers to deliver (and sometimes plan) instruction, (b) lessons were, or were well situated, within a regular school curriculum, and (c) the assessment methods examined outcomes in standardised tests, sometimes alongside bespoke measures aligned to the specific learning content. Ecologically valid studies were designed to be sustained over time; they were tested at scale with many teachers and across many settings, and the approach was designed to have wide applicability across one or more subject areas.

The majority of studies we have reviewed, however, have not met this description. Studies of applied cognitive science that are both rigorous and ecologically valid are, at present, all too rare. Our ecological validity screening tool and reporting (that is, each strategy evidence review begins with a summary of the main characteristics of the data) retains these questions in focus and, in our view, enables a degree of theoretical generalisation from studies that, while they have some contrived elements, nonetheless present evidence that is not overly distant from practice.

What is the role of teachers in applied education science? What are the challenges of working with teachers and schools outside of laboratory settings or tightly controlled studies in schools?

Another group of studies that is represented in this section by Feddern et al. (2018) are those that make use of independent study or computer applications to deliver the intervention. There is huge merit in considering the self-study and technology-based interventions. For our present purposes, however—which centre on concluding about whether evidence supports the use of cognitive science principles within classroom practice—designs that minimise or bypass teacher input raise questions such as those we pose below about teacher learning and implementation. Feddern et al. (2018), in this section, provides a good illustration of this group of studies. Their intervention focused on using a computer programme to support revision, comparing this to a PDF or physical revision guide. This was a well-conducted trial, clearly testing the cognitive science principles and doing so in a way that could significantly improve student results. The quality of the study, in this respect, clearly highlights the distinction between concluding that (a) a cognitive science principle holds true, that (b) a cognitive science principle can be applied in a school setting to impact curricular learning, and (c) that a cognitive science principle could and should be adopted by teachers in general across their practice. Computer-based or independent study approaches—by cutting out the teacher—are in a position to test b, and potentially a, but not c. It is also interesting to consider the converse: some of the studies in this mixed strategy section have been strong tests of wider adoption by teachers, but—as a greater number of strategies and practical factors are introduced—in the 'open-system' of the education system, it becomes hard to draw conclusions about specific strategies.

A related issue we have encountered is that even where teachers are involved, where these were not manifestly representative of the wider population of teachers—as was usually the case—there is a question about to what extent an intervention is a test of the quality of teaching rather than the cognitive science principles at play. For argument's sake, suppose that the postdoctoral and doctoral researchers delivering the intervention in Barbieri et al. (2019) are highly skilled classroom teachers and that the positive result was caused by, or at least dependent, on this fact: it is one thing to find that an expert or research-enthusiastic teacher can make a strategy work and another to make generalisable claims about the suitability of a strategy for the wider teacher population or curriculum. Again, while a small degree of theoretical generalisation is defensible (and inevitable), strong conclusions can only rest on studies that are conducted at sufficient scale with teacher samples that are representative of the wider teacher population targeted.

We bring this general discussion of translation and ecological validity to a close by relaying an account of the challenges from Davenport et al. (2020), one of the studies in this mixed strategy programme area. This study focuses on the problem of applying findings from cognitive science in 'authentic settings', something they observe is challenging and a gap in the literature. They describe the set of challenges as follows:

> *Translating findings from the lab into effective classroom instruction is not straightforward. Many laboratory studies narrowly focus on learning principles in isolation, use populations of undergraduates, and study interventions over short periods of time. In contrast, classroom-based instruction requires simultaneously integrating and applying a variety of learning principles in a complex and dynamic system that involves teachers and entire classrooms of students over months and years (p.516).*
>
> *[…]*
>
> *Design principles derived from research range from general and abstract to specific and directive (Cremers, Wals, Wesselink, & Mulder, 2017). Practitioners often struggle to interpret and integrate abstract, general principles with everyday instruction (Coburn* et al., *2009; Kochanek, Scholz, & Garcia, 2015) and similarly struggle to transfer highly specific design principles, illustrated in narrow educational contexts, to other situations (Kali, 2006). Individual research principles offer little guidance about how they should be used to teach specific content, how often they should be used, or when they should be used (e.g., at the beginning versus the end of a school year). Further complexity arises as strategies are combined […] As a result, design principles are "necessary but not sufficient for the production of excellent tools for practitioners" (Burkhardt, 2006, p. 132). As instructional design is ultimately a form of engineering, moving research into practice requires that theories are integrated and applied with a focus on practical impact and evaluation at scale (Burkhardt, 2006; Kirschner, Verschaffel, Star, & Van Dooren, 2017). (p.517)*
>
> (Davenport et al., 2020)

The gaps between findings and learning principles and classroom practice are also discussed in Schunn et al. (2018, p.238); they discuss the 'question of translation or operationalization'. Their work was focused on curriculum materials for which 'there were many possible ways principles could be implemented'. As they note, their application to two whole-curricular rather than discrete topics makes it less likely that their 'results depend upon one particular implementation'. While variations in implementation may well 'wash out' in overall effect estimates across a larger and sufficiently representative sample, it does raise the questions of (a) whether 'tighter' or more accurate principles and a greater fidelity to them during implementation is one way in which studies such as those considered here could successfully apply cognitive science to the classroom and (b) to what extent we can and should specify the necessary principles for classroom success. We are reminded of Wiliam's (2016) discussion of 'tight by loose' frameworks:

> *The idea behind the tight but loose framework is that it provides a way of managing a delicate balance act between two conflicting requirements. The first is the need to ensure that the model is sufficiently flexible to allow it to be adapted to the local*

> *circumstances of the intervention not only just to allow it to succeed, but also so that it can capitalise on any affordances present in the local context that will enhance the effectiveness of the intervention. The second is to design the model with sufficient rigidity to ensure that any modifications that do take place preserve sufficient fidelity to the original design to provide a reasonable assurance that the intervention does not undergo a 'lethal mutation' [i.e., an adaptation that violates a vital operative aspect of the intervention and reduces its effectiveness].*
>
> (Wiliam, 2016, p.209)

A difficulty with the present level of understanding of cognitive science in the classroom is that the principles we have are based on research and not typically tested in the open-system conditions of the classroom. We have endeavoured to set out the conceptual landscape of these principles drawing on the wider literature as well as our own wider evidence in the study database. Much of the wider literature is based on those actively applying cognitive science in classroom contexts. In the next section, we present our practice review and results from surveys exploring teacher perspectives on the theory and practice of applied cognitive science. These efforts notwithstanding, and with particular admiration of the creativity and expertise of educators currently developing how, for example, to apply dual coding across all areas of the curriculum (Caviglioli, 2019), we are not in a position where we can match the practice with rigorous and authentic classroom trials.

Professional development and learning

What is the role of professional development and learning for applied cognitive science?

The remainder of this section briefly discusses more general considerations and principles that will be important for larger scale implications.

Many of the studies in this area have combined curriculum (re)design with professional development for teachers to apply cognitive science principles at scale. There are certain advantages to using the curriculum as a key vehicle for this—in particular, preventing the need to teachers to have to re-create plans and resources to implement cognitive science strategies. This also applies on a smaller scale with schemes of work and lesson plans. As well as saving time, focusing on curriculum design and the provision of resources can 'capture' principles in practical tools and promote successful implementation. As above, getting these 'tight but loose' is likely to be a key design factor for this. One of our studies in this area, Yang et al., (2020), encountered difficulties with implementation; they reflect that providing materials was insufficient for transforming instruction, adding:

> *The Cog-sci teachers might have benefited more if our professional development (PD) had offered more direct experiences with the optimal learning environment they are expected to construct (Marx & Harris, 2006). Another explanation is that average effects may be masking substantial individual variation. Previous research has shown that what teachers learn in PD depends largely on the existing knowledge they bring to the activity and that they can have quite different takeaways from their learning experiences (Minor, Desimone, Lee, & Hochberg, 2016). Personalizing PD to address teachers' particular circumstances, knowledge, and experience holds promise for increasing their effectiveness (Starkey* et al., *2009).*
>
> *(Yang* et al., *2020, p.558)*

Yang et al. (2020) also drew attention to the distinction between content knowledge (in their intervention, teachers were taught science content at introductory undergraduate level) and pedagogical content knowledge, 'which incorporates knowledge of how to model and illustrate concepts and how students learn in a particular content area' (p.541). What is clear within this explanation, and the literature more generally, is that even when providing curricular materials for use, professional development and learning are likely to be vital aspects of successful implementation. As Fullan (1992, p.23) explains, 'If implementation involves new behaviours and beliefs, teacher development in relation to these new learnings is a *sin qua non*. This is why in-service and professional development in support of specific interventions is usually found to be the critical factor for success (see Huberman & Miles, 1984).'

Therefore, one likely strand of a theory of change for an ecologically valid and large-scale cognitive science programme will relate to effective professional development (Cordingley et al., 2015; Darling-Hammond, 2017).[28] Looking at the studies in this mixed programme section, Desimone and Hill's (2017) discussion of the implementation of Schunn et al. (2018) provides helpful detail on a programme that incorporated a strong professional learning aspect. While the study results were disappointing (see above), the implementation was thought to be successful. Desimone and Hill (2017) describe the professional learning approach within the programme as follows:

> *For both treatment conditions, PD was taught by science museum professionals, university professors, science researchers, and high school content area teachers who specialized in the given content unit. Follow-up sessions during the academic year were modelled as professional learning communities (PLCs) for both treatment conditions—CS and Content. The PLCs gave teachers an opportunity to share their successes and difficulties with instruction and offer guidance and support to one another. For the CS condition, curricular modifications continued to be presented as part of the PLCs.*
>
> *All PD provided as part of the intervention reflected five key features of high-quality PD that have been shown in rigorous empirical studies to be related to changes in instruction: PD was focused on content, included active learning opportunities for teachers, was coherently integrated into the curriculum, provided a substantial number of sustained contact hours, and included collective participation of teachers from the same subject (e.g., Desimone & Garet, 2015; Garet, Porter, Desimone, Birman, & Yoon, 2001; Penuel* et al., *2007). Specifically, teachers in the CS and Content treatment conditions participated in 2.5 days of summer PD for each unit prior to the unit beginning (see Figure 1 for our implementation timeline). Teachers also participated in four follow-up PLCs during the semester in which they implemented a given unit; these lasted 2 hour each. In total, for each unit they participated in, in the first year of the study, teachers could receive up to 18 hours of summer PD and 8 hours of PLCs (2 hours per PLC meeting; PLC meetings took place on a monthly basis for 4 months).*
>
> *(Desimone and Hill, 2017, p.515–16)*

This focus on professional development is evident in most of the studies in this area, albeit described in less detail than Desimone and Hill's dedicated implementation study. There were some links with

[28] An EEF review of the evidence on CPD is also forthcoming.

specific cognitive science principles evident within this, such as Davenport et al.'s (2020) discussion of worked examples and spacing:

> *Some principles required more changes to teacher practice than others—for example, implementing the worked-example principle relied primarily on the use of revised student books, whereas implementing the spacing and formative assessment principles relied heavily on teacher practice.*
>
> *(Davenport et al., 2020, p.523)*

Practical challenges for implementation and evaluation

What are the challenges of implementing mixed cognitive strategy programmes?

Many studies in this section discussed challenges relating to implementation. Schunn et al. (2018) discuss several common issues that can arise:

- teachers may lack the autonomy and control required to make some targeted changes (for example, there may be school leadership or policy requirements to follow);
- student attendance may be low, diluting the impact of the intervention; and
- socioeconomic factors or challenges stemming from pupils' home background can reduce the effectiveness of instruction.

They also discuss the effect of multiple factors and their concentration in individual schools:

> *Disruptive factors co-occur in complex combinations and can occur at many levels—either specific to the classroom (e.g., a group of students are collectively unruly), to the teacher (e.g., the teacher has poor classroom management skills), or to the school (e.g., extracurricular announcements regularly take priority over quiet classroom time).*
>
> *(Schunn et al., 2018, p.228)*

Schunn et al. (2018) was a study that saw the programme implemented with high fidelity (p.233). Other studies had less success in this respect. Davenport et al. (2020) reported significant issues with attrition, despite teachers receiving a stipend for participation. The majority of this stemmed from factors from within the schools such as teaching reassignments and curriculum changes. More widely, there were issues relating to the implementation of the curriculum. The study collected logs from teachers recording which curriculum units had been successfully implemented. Implementation of curriculum units ranged from 30% to 61% in control schools and 36% to 65% in treatment schools: 'No schools implemented the entire curriculum and its components as designed, completing at most two-thirds of the curriculum' (p.527), and 'nearly 90% of participating schools did not implement the eighth unit' (p.528). Such shortcomings in implementation are common: similar issues arose in Yang et al. (2020). It is a strength of these studies that these issues are made visible and accounted for in the analysis and discussion. Getting cognitive science into practice involves not only getting the learning principles right, but also the operational, practical aspects of organisational change. This latter set of issues need to be treated seriously in their own right. We are in strong agreement with a point Davenport et al. (2020) make in their conclusion:

> *Effective instructional design is ultimately a feat of engineering rather than a natural consequence of scientific research on learning. Instructional designers must*

> *integrate information from many sources as they create new materials. The process described in this article offers concrete steps that articulate how to overcome some of the challenges that arise in translating research into practice and demonstrates how controlled experiments can be used to verify that design iterations result in improved outcomes.*
>
> *(Davenport et al., 2020, p.531)*

Understanding implementation

What was found out about implementation by Desimone and Hill's (2017) focused study of the implementation of a cognitive science intervention?

We close this section with a summary of key findings from Desimone and Hill (2017). This study was the only study in this section specifically focused on implementation issues, which are likely to be applicable for all school improvement focused studies in this overall review. Implementation fidelity refers to the degree to which an intervention or programme is delivered as intended. As Desimone and Hill note, research into implementation fidelity is currently severely lacking. As they explain, implementation research is not merely a case of looking to appraise the main results in an effectiveness trial, but rather an integral aspect of an evaluation that can yield important results about the operative components of more complex interventions and understand how an intervention was adapted to and interacts with the contexts and purposes to which it was applied. They reach several key conclusions on what an implementation study can reveal:

- **An intervention may work partly through secondary mechanisms**. They found that the structure and sequencing of the intervention supported teachers' classroom management and 'provided a more coherent organisational structure to their daily lessons' (p.528). These benefits were not central to the main objectives relating to cognitive science principles but may have been a factor in improving teacher practice.
- **An intervention's effectiveness may be related to balancing teacher content knowledge, aligned lesson plans, and teacher invention**. Desimone and Hill (2017) present an interesting analysis unpicking (through a structural equation model, Figure 3, p.528) the direct and indirect influences of teacher subject knowledge, cognitive science principles, implementation factors, and background teacher and student factors. Their comments align with Wiliam's 'tight but loose', as described above.

> *We found that there was no significant relationship between content knowledge and implementation, and furthermore that teachers with higher content knowledge did not implement the intervention more frequently or better […] This raises a question about trade-offs between a scripted intervention and one that requires considerable teacher knowledge and invention […] We suspect that the success of the intervention was partly due to the balance of research-based approaches (i.e., applying CS principles to teaching), [professional development] that included both content and pedagogy, and implementation that provided aligned-lesson guidance while still allowing for teacher creativity and invention.*
>
> *(Desimone and Hill, 2017, p.529)*

- **Professional learning communities (PLCs) help teachers refine and adapt an intervention**. The professional learning that accompanied the intervention was held to be of great value,

with peers providing support to implement the materials and principles in practice. Desimone and Hill describe the 'trial and error efforts, to find out what was working, and to share ideas and experiences with other teachers to improve their implementation of the CS principles' (p.529).

This discussion connects to principles from the wider school improvement literature. Yang (2020), for example, refers to principles such as the 'implementation dip' (Fullan and Miles, 1992) where changing practice can reduce effectiveness in the short term as the teachers become accustomed to the new approach. They discuss the differential effectiveness of interventions for different student groups and abilities, and issues with cognitive science materials being suitable for some students, but not ('too detailed', p.558) for others. As well as being a general principle of ensuring materials are suitable for students, this brings us back to specific cognitive science principles we have discussed, in particularly around the management of cognitive load, for example. Given the expert reversal effect, a cognitive load management intervention would need to be tailored (perhaps through suitable in-built formative assessments) for specific groups of learners. This is but one example: the more general point, well made by Desimone and Hill (2017) is that '**implementation matters, and ought to be measured**':

> *Although research has demonstrated that high levels of implementation fidelity translate into improved student outcomes (Durlak, 2010; Durlak and DuPre, 2008; Kaderavek and Justice, 2010; Stein* et al., *2008), in some cases, teachers may not be implementing the intervention due to contextual or environmental pressures, lack of knowledge, or any one of a myriad of other factors. Before concluding that an intervention does not have effects on student learning because RCT results show no effects, it is imperative that researchers have a measure of whether or not the teachers actually implemented the intervention as intended.*
>
> *(Desimone and Hill, 2017, p.527–28)*

We very much agree. We advocate implementation (or process) study that 'looks in', examining the internal logic model (or theory of change) for interventions and 'looks out' at the contextual factors that influence a programme's implementation and effects. For example, Ainscow, Chapman and Hadfield (2019) devote a chapter to addressing barriers to change, including discussing the social, political, and cultural factors that can influence whether school improvement programmes are successful. Connecting to, and evaluating, these 'inward' and 'outward' programme effectiveness factors has value when assessing effectiveness, whether results are positive or not. The import of this for this particular review is that many of the programmes we have reviewed have not offered sufficient evidence to assess programmes on these multiple levels; furthermore, the evidence-base—by not always containing authentic applications of cognitive science to authentic classroom environments taught be regular teachers—makes a comprehensive assessment of our focus strategies challenging, even when drawing from across the combined evidence-base.

Final thoughts on this strategy area

Ecological validity has been a key concept across this review. It has been a key distinction between applied and basic science, and a question which lies right at the heart of this review. We have not, beyond some isolated points, questioned the validity of the basic cognitive science evidence-base. Most of the principles we have focused on are well known and are associated with a large basic science evidence-base going back many decades. While of course this remains contestable and the basic science continues to advance, whether or not the basic science holds has not been central to our investigation. The position that *has* been central is that we cannot take it for granted that basic

scientific results established through laboratory studies or conditions lacking ecological validity will necessarily work in the classroom. The focus of this area has been mixed strategy programmes. As we have described, several of these studies have had high ecological validity and were some of the strongest pieces in this respect in our database. We, therefore, evaluate these results with both the potential of mixed strategy programmes and questions—about applied education science more generally—in mind.

In our systematic review of classroom trials, we concluded that evidence provides either positive or mixed/neutral results for programmes of mixed cognitive science strategies. Our confidence in this was low. The evidence showed that, at present, there are few or no large-scale mixed strategy cognitive science programmes that have been trialled and found to be effective. Of those that have been trialled—of which there are only a small handful of ecologically valid, rigorous examples—several have yielded disappointing results.

These results led into a discussion of implementation. Testing interventions in realistic settings under realistic conditions provides a stronger warrant for recommending that effective strategies are adopted more widely. It does, however, make isolating the principles at play more challenging and the results more complex, and less certain. We return to this question, first, in the summary of (non-strategy-specific) perspectives of practitioners in our Practice Review section and, second, in the context of discussing the overall implications of the study.

B9. Practice review perspectives

Introduction

As outlined in Part A, the practice review had two main objectives and two main components:

The overall objectives of this practice review are to answer the questions:

1. What applications of cognitive science in the classroom are currently prominent in policy, guidance, and practice? What do practitioners in England identify and recognise as common approaches based on cognitive science?
2. What form(s) do applications of cognitive science take when manifested in practice? How do cognitive science applications differ for different contexts, subjects, and groups of students?

In overview, it comprised two main activities:[29]

- **A literature review**. Alongside the main review, we reviewed literature to identify applications of cognitive science in the classroom from policy and practice documents (for example, reports, frameworks, guidance, and popular-scientific texts). The bibliography used for scoping and protocol development is provided in Appendix 2. The final bibliography, developed and accessed throughout the review, is provided in Appendix 13.
- **Data gathering, screening, and extraction**. We used interviews and a questionnaire to survey practitioners in England. Questions were developed as part of the practice review based on the questions above and refined following mid-point analysis from the core systematic review. Our survey was distributed via teacher and school organisations and social media.

The practice review strand of this study has run alongside the review. However, most of its impact can be seen in connection with reporting in other areas. In particular, it:

- laid the groundwork for the review protocol (see bibliography for scoping and cognitive science concept map);
- informed the categorisation of studies across the review areas and strategies; and
- informed and provided data for the 'Evidence-informed discussion and questions' sections following each of the main evidence reviews via:
    - identifying key concepts, moderators, and implementation factors,
    - locating practitioner-focused accounts of cognitive science, and
    - providing practitioner perspectives from interviews and questionnaires on the areas we have reviewed.

Therefore, most of the practice review output and outcomes have formed and been reported within other sections with this report. In particular, perspectives relating to the focus strategies have been integrated into the relevant review areas. There was, however, more general data arising from the practice review that did not connect to the specific strategy sections. Notably this included perspectives on cognitive science in general in the practice-focused literature and in our interviews and questionnaire. Many teachers and commentators provided general points about the current state of the art, the value, the future, and the challenges posed by cognitive science in the classroom.

[29] We include full details of methods for the practice review in Appendix 13.

This section reports, first, general points from the practice review literature and, second, perspectives from the practice review interview and questionnaire data that have not been reported elsewhere in this report. We have not done a systematic review of the practice-facing literature and the teacher perspectives we present are *not* representative of teachers more widely, given the self-selected nature of the sample. Moreover, we are not presenting teacher perspectives as views we either agree or disagree with. Our aim is to air a range of perspectives on cognitive science and summarise and list, rather than comment on or analyse, our data.

Practice review literature

On making connections between basic science, applied science and education research, and educational practice.

In schools, there is a hunger for more knowledge about the brain as well as a concern that teachers are sometimes being provided with a range of 'neuromyths' (Goswami, 2006). In the literature on cognitive science in education, the two disciplines—neuroscience and cognitive science—are commonly described as being characterised by a 'gap' (Howard-Jones, 2014) in need of 'bridging' (Aronson, 2020) or even a 'bridge too far' (Dougherty and Robey, 2018). Dougherty and Robey (2018), for example, argue that neuroscience is largely useless for education without including cognitive and behavioural science as a 'middle-man'.

Churches et al. (2020) describe three key challenges of collaborations between neuroscientists and educators: (1) the two disciplines are fundamentally different in their objectives, (2) research within the two disciplines can take place at different levels and meet at the behavioural level, and (3) the problem of translating neuroscience research into something applicable in the classroom and following a school timetable. For example, they argue in relation to retrieval practice that 'particularly in the form of multiple-choice testing, [retrieval practice] is often operationalized in a way that would be of little direct benefit in the classroom over a whole 1-hr lesson period' (p.6). A whole lesson includes many other elements, such as interactions, feedback, and different types of instruction and sometimes tests may run counter to other effective classroom strategies.

There are also connections between challenges faced by researchers and by teachers. In many of the review areas, we have discussed the difficulties teachers have faced in understanding and implementing the strategies. This is also reflected in the research context. Weinstein et al. (2018) say that, 'Future research needs to (a) better formalize the definition of each strategy (particularly critical for elaboration and dual coding), (b) identify best practices for implementation in the classroom, (c) delineate the boundary conditions of each strategy, and (d) strategically investigate interactions between the six strategies we outlined in this manuscript' (p.13).

Despite these challenges, there were numerous points made about the value of creating a common language and promoting two-way dialogue. Kelleher and Whitman (2018) discuss the value of a common language as follows: 'One added benefit of using a Mind, Brain and Education research lens for this work is that the common language and research bases helps teachers in different disciplines and divisions have fruitful conversations on a common goal' (p.228).

On, applicability, effectiveness and practice variation.

The literature revealed a diverse and disparate picture of how cognitive science is used and understood in the classroom. Not only did sources identify a range of strategies, they also illustrated

that each practice encompasses significant diversity in relation to how it is practiced and, furthermore, is supported by other supplementary strategies (for example, feedback; Dunlovsky and Rawson, 2015). Moreover, many observed that strategies are not practiced in a vacuum: Darling-Hammond (2020), for example, argues that the success of instructional strategies and learning more broadly depends on, for example, safe environments, supportive relations, avoiding stereotypes, and socio-emotional learning. One example of this was Fazio's (2019) study of retrieval questions within middle school mathematics classrooms. Fazio (2019) argues that in order for retrieval to benefit students, they need to be given time to respond and there needs to be a culture of participation in the classroom.

There was some discussion of applicability and take-up relating to country contexts. Most of the literature included in this review comes from high-income countries and, given the aim to inform teaching practice and debates in England, this is also where our focus has been. As argued by Abdazi (2014), it is important not to assume that strategies work equally well or are equally suitable in low-income countries and therefore it is important to emphasise that they cannot be seen as universally applicable. Cognitive science is not equally embraced everywhere (Aronson, 2020). Even within high-income countries such as the U.K., educational settings vary significantly with regards to the student intake and school resources and there may be differences with regards to the specific subjects being taught.

Most warn against blanket judgements of effectiveness or applicability. Churches et al. (2020), for example, conclude that, 'It cannot be good enough to imply that testing will always work, for every teacher, in every situation, with all children—nor can it be acceptable to jump to similar conclusions about other evidence from the science of learning' (p. 6). Similarly, Dunlovsky and Rawson argued that there is still much to be discovered about how to best take advantage of these techniques in the classroom as much of the evidence is from laboratories. However, they argue that 'this limitation in our knowledge [is] an opportunity, because as we all are trying out these techniques, we can conduct relatively straightforward investigations to evaluate their efficacy' (p. 77). Similarly, Howard-Jones (2018) argues that, 'Although the science provides principles and a scientifically determined understanding of how learning works, based on concrete measurement of behaviour and brain function, it does not provide a list of 'top tips' or practices that are guaranteed to work with any class or individual in any context. In the absence of a one-size-fits-all prescription for effective teaching, teachers must constantly make decisions based on their own ideas of how learning proceeds and what they observe occurring in their classrooms' (n.p.). Kelleher and Whitman (2018) mention the importance of letting teachers choose and then see how different strategies work for their classrooms, and that it evolves over time. Our interviews also had a time component with teachers talking about experimentation and not always getting it right.

Notwithstanding the recognition that general applicability and effectiveness is unlikely, many have sought to probe the boundaries of this, looking for strategies that work across subjects and for most children. Dunlovsky and Rawson (2015) argue that: 'Of course, an all-purpose technique that will solve every problem that struggling students have is not currently available, and we suspect it never will be, because even the most versatile techniques have limitations. Nevertheless, several low-cost techniques have demonstrated generality in their effects on student learning and can be widely applied' (p. 72).

On implementation

Many of the cognitive science strategies may take time for teachers to get to grips with for best effect. One consideration in implementation is therefore professional development and learning and its role for the implementation of cognitive science informed practice. Howard-Jones (2014) mentions the lack of training as an issue: 'Surveys of teachers in countries with very different cultures have revealed similarly high levels of belief in several neuromyths. This prevalence may reflect the fact that neuroscience is rarely included in the training of teachers, who are therefore ill-prepared to be critical of ideas and educational programmes that claim a neuroscientific basis' (p.817). Howard-Jones (2014) also mentions that neuromyths may be allowed to thrive because potential counter-evidence is presented in technical neuroscientific journals that are not accessible to teachers (also see Kelleher and Whitman, 2018).

Purdy (2008) says that a process of supporting teacher development will necessitate more effective dissemination of the most recent neuroscientific research findings to the educational community (p.204). However, the interviews also showed a wish from teachers for practical, hands-on training, observations, and communities of practice. Teachers wish to see how it works 'in practice'. These are beyond the scope of this research, but a clear area for potential further research.

This review has not included literature on meta-cognition, although this is strongly related to our focus and of interest to many teachers and professionals (for example, Carey, 2014 and examples in our interviews). Related, Dunlosvky and Rawson (2015) talk about the importance of teaching students learning techniques. Implementation of cognitive science informed practice would require consideration of the understanding and ownership of learning approaches by students as well as teachers. We have certainly been conscious in this review of the fact that most of the focus has been on the teacher selection and implementation of strategies. There were, however, many studies in which the approach to learning was highly student-led, and in particular our discussion of generative learning (Section B5) emphasises the potential value of this.

Implementation was also considered at a larger level, with discussion of faculties, schools, and systems and their role in implementation. For example, Whitman (2018) talked about an 'all in' model where all of the faculty must be trained in and expected to use research to inform their practice' (p.3) and argued that this was pivotal to the success of their Mind, Brain and Education programme. We also have noted (in Part A) the prominence of cognitive science in the documentation and policy of Ofsted, the schools' inspectorate, the National Professional Qualifications frameworks, and the Early Career Framework training programmes.

At a teacher level, teachers frequently mentioned time as a limitation in developing material using these strategies. Many challenges seem to be about time constraints and curricular constraints. This relates to Kelleher's and Whitman's (2018) view that strategies have to be worth the time and effort because it will take time for teachers to implement new practices.

Practitioner perspectives on cognitive science

In this section, we report general results from our practice review questionnaire and interviews. We stress that the questionnaire sample is not a representative sample of teachers and is skewed towards having more secondary teachers, teachers with more experience, and—as far as we can tell—teachers more positively disposed towards cognitive science than typical teachers. A range of perspectives are aired, however, including both positive and negative views. Our intention here is to outline this range of perspectives.

Questionnaire responses

How would you rate your knowledge of these cognitive science strategies and how they apply in the classroom?

	Not heard of	Low knowledge	Medium knowledge	High knowledge	Total
Spaced practice	6.1%	10.2%	39.5%	44.3%	512
Interleaving	6.8%	16.0%	41.3%	35.9%	513
Retrieval practice	1.2%	2.8%	26.9%	69.2%	509
Dual coding/multimedia learning	3.5%	14.6%	43.8%	38.1%	514
Strategies to manage cognitive load	1.4%	9.6%	44.6%	44.4%	511

In summary, teachers responding to our survey tended to have:

- higher knowledge of retrieval practice; and
- lower knowledge of interleaving and spaced practice.

Have you completed any professional development and learning in these (select any that apply, leave blank if you have not completed any)?

	Training provided by own school/trust	Training provided externally	Independent learning	Collaborative/ peer learning with colleagues or networks	None	Total
Spaced practice	20.3%	11.4%	39.2%	17.1%	12.1%	780
Interleaving	17.7%	10.9%	39.8%	15.6%	16.1%	719
Retrieval practice	24.2%	13.1%	37.9%	19.6%	5.2%	854
Dual coding/ multimedia learning	18.4%	13.4%	39.3%	16.8%	12.1%	745
Strategies to manage cognitive load	24.1%	11.2%	40.4%	17.6%	6.7%	775

In summary:

- Many teachers responding to our survey had pursued independent learning about cognitive science. We expect that there is a strong self-selection bias in the survey sample for teachers who independently learn about cognitive science being more likely to respond to our survey.
- Sample bias notwithstanding, the other results suggest that there is applied cognitive science training available but even amongst our sample, many had not received formal training.

How important do you think these strategies are for effective teaching and learning?

	Not important	Low importance	Moderate importance	High importance	I don't know	Total
Spaced practice	0.4%	1.0%	18.1%	73.7%	6.9%	509
Interleaving	0.6%	3.6%	33.0%	51.7%	11.1%	503
Retrieval practice	0.4%	0.6%	7.1%	89.9%	2.0%	505
Dual coding/ multimedia learning	0.6%	4.7%	34.4%	52.1%	8.3%	509
Strategies to manage cognitive load	0.4%	0.6%	14.8%	81.9%	2.4%	508

In summary:

- Spaced practice, retrieval practice, and 'strategies to manage cognitive load' were identified as most important for effective teaching by survey respondents.
- Interleaving and dual coding/multimedia learning were rated as relatively lower effectiveness but still with the vast majority of teachers stating that these were of moderate to high importance.

Please state your level of agreement with the following statements:

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree	Total
Learning about cognitive science strategies has improved my teaching	60.4%	31.9%	4.7%	0.6%	2.4%	492
Cognitive science is a new way of talking about old teaching strategies	10.9%	41.9%	21.6%	17.5%	8.2%	487
All teachers should be taught cognitive science informed teaching strategies	71.1%	23.4%	4.0%	0.6%	1.0%	505
Cognitive science informed strategies are central to my own approach to teaching	51.6%	35.0%	8.8%	3.2%	1.4%	500
I think there is firm scientific evidence to support all or most of the cognitive science strategies (as above)	44.7%	41.3%	10.9%	1.9%	1.3%	479
I find it difficult to implement cognitive science strategies in the classroom	2.4%	19.2%	13.7%	32.5%	32.1%	495

In summary:

- Teachers responding to our survey view cognitive science strategies as effective, for them and more generally.
- Many believe that cognitive science is a new way of talking about old teaching strategies.
- Most believe that there is firm scientific evidence to support all or most of our focus strategies.

Please state your level of agreement with the following statements:

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree	Total
In practice, I find that a lot of the cognitive science strategies overlap with one another	19.3%	66.0%	10.1%	4.0%	0.6%	497
Students should be taught cognitive science informed strategies	45.8%	39.5%	11.9%	2.0%	0.8%	494
Cognitive science informed strategies work best when implemented consistently throughout a school	61.1%	28.4%	8.2%	1.9%	0.4%	486
A teacher should have the freedom to select which cognitive informed strategies to implement in their own classrooms	34.7%	41.7%	9.3%	11.9%	2.4%	504
Using cognitive science informed strategies is a useful way to provide support for pupils with additional learning needs	56.3%	35.3%	6.8%	1.0%	0.6%	485

In summary:

- Many teachers responding to our survey believe there is overlap between the strategies.
- They believe that the strategies are effective when they are implemented widely across the school, including (with high but slightly lower agreement) teaching them to students.
- They believe that the strategies are valuable for supporting pupils with additional learning needs.
- Most strongly agree or somewhat agree that teachers should have the freedom to select which cognitive informed strategies to implement in their own classrooms.

General interview responses

Below, we report the perspectives of teachers responding to our interviews and open-response sections of our questionnaire. We have organised these comments around several general themes. We pose a question at the top of each sub-section followed by a selection of quotations. We keep our comments and analysis to a minimum. As described above, we present these quotations as a range of perspectives on cognitive science. We are not endorsing these views, nor have we been selective in what we have and have not reported. We have simply grouped and summarised a range of general teacher perspectives on applied cognitive science to provide a picture of the variation of perspectives of teachers in our sample.

Applicability

Does cognitive science equally apply for all age groups?

In our questionnaire we asked teachers whether there were ages, pupil groups, or subjects for which cognitive science is more or less applicable. Many used this question as an opportunity to comment about the strategies being generally applicable; some commented on groups or subjects with which they have had more or less success.

Several respondents expressed the view that cognitive science strategies are suitable for all age groups:

- *'They are effective strategies that work well across all key stages.'*
- *'They all seem to work very well for ALL ages and ALL groups.'*
- *'All pupils benefit from these approaches.'*
- *'No, I was surprised to note that the same strategies worked for my year one class as they do for my year fives.'*
- *'I think, in general, many primary educators feel cognitive science is too difficult for young students and I strongly disagree. I think if we introduce these concepts and teach students how to learn and assess themselves at an early age, we are empowering them to drive their own learning in middle and secondary school.'*

Several discussed groups for which cognitive science strategies were particularly applicable, many making a link to exam preparation:

- *'It has been effective in all age groups but particularly useful at KS4 and KS5 ... works well for exam groups as they are motivated and invested in wanting to retain large amounts of knowledge.'*
- *'They are much more applicable to exam groups, who need to recall facts, rather than for the more practical project-based work that KS3 do. In my school, students choose GCSE options from Year 9, so they become more important in my teaching from that point.'*
- *'I think they are useful to all students. I think some students have a better ability to self-regulate. When we teach them things like revision, spaced practice … they might get that really quickly or might already be doing it unconsciously. For others it is brand new. And it takes a bit more of an effort to make them work that well. The students who are self-motivated and have it from home, they are useful, but they would probably get there on their own anyway. It has a bigger impact teaching it to students who wouldn't otherwise come across them.'*
- *'Older students (KS4) often have preconceived ideas that can make using these strategies more difficult. For example, dual coding and retrieval practice for revision are more useful but some students still fall back on reading their notes (a low utility approach).'*

Subjects

Does cognitive science equally apply for all subjects? Are you applying it?

In general, respondents thought that cognitive science strategies were applicable across subjects but that there were important differences in language, applicability, and practice, and some dangers in not considering these:

- *'Pedagogy does need to be flexible within different subjects. And that is something that I think senior leaders don't always grasp. So, for example, within PE, the way in which we would need to deliver might differ from how you would deliver English in the classroom. And it's … it's understanding that you need a certain degree of flexibility, whilst those core tenants need to be there to be research informed. I found it quite difficult in the first instance to find cog sci in PE and then I discovered it was within different terminology. When we talk about spaced practice or interleaving, they talk about variable practice. It is the same idea with a different name' (Interviewee 10).*
- *'I have found it is very effective in maths and foundation subjects, which are predominantly knowledge-led subjects. Within English, it is effective when teaching SPaG [spelling, punctuation, and grammar] and guided reading, however they are harder to implement when teaching writing. In order for pupils to produce a high-quality piece of writing, they need to apply such a broad range of skills as well as take part in high quality discussion to generate ideas; strategies such as quizzing are less effective. We still ensure that when planning SPaG and guided reading that the skills they learn and use in these sessions feed into their writing lessons, which aligns with spaced learning and retrieval practice.'*
- *'With regards to whether cognitive science strategies work best when implemented consistently: I think broad principles should be consistent and explained to students but that the way they are used would need to be different for different subjects.'*

One interviewee discussed the role of a whole-school model being contextualised by subjects to strike the balance between consistent approaches and subject-specificity (see below).

- *'I think there's a tendency in arts-related subjects such as art drama or music, to think, "Oh, well, this doesn't apply." But actually, it does. And it's more a case of helping those subjects to understand what the thing is, and then how it might fit into their context. That's why, from a whole-school point of view, you've got to set your model. And you've got to set the course of direction, as opposed to a ground-up approach where subjects are left to work it out for themselves. And eventually, over time it spreads across the subjects. I think what we do that's quite different is that when we want to do something, we learn how to do it. And we learn that we then do it as a whole school, which later is then contextualised into subjects once we've worked out what, or understood what it is we're trying to do. And then we also have a culture of learning over time. So we know that we might launch a thing. But that is just the start of the journey. And two, three years later, we're still trying to get it right, whatever it is' (Interviewee 13).*

One questionnaire respondent felt that the applicability should be judged by assessment of pupils and teachers need:

- *'In general, though, I'd say that nothing works in a classroom without some diagnosis of what the pupils and teachers need in that particular setting.'*

Implementation

What are the main barriers to implementing cognitive science strategies?

A common issue mentioned regarding the implementation of cognitive science strategies was time:

> *'But the thing is, it takes lots of time and it is not as fast as I would like it to be, to make sure it is embedded. Because teachers are so busy and it is a very demanding job, responding to the emotional needs of people all the time, and in the context of*

> *the pandemic, that whole space for teaching and learning—teachers are kind of in survival mode. It is only really from the second half term we have actually been talking about teaching and learning again … we are only picking a couple of things, because it takes a while to change habits and it is a whole-school thing that we want to do, but staff have got that time as well to build things up and take risks. Learning is a complicated process, isn't it?'*
>
> *(Interviewee 1)*

Similarly:

> *'Teachers are time-starved, so you have to drip feed; it can't be massive changes to how they teach. It's a slow change. You have to prioritise little things that make a change now. That's the biggest barrier' (Interviewee 6).*

A second, large group of comments relating to common barriers to implementation were issues of securing 'buy-in' and understanding from teachers. Some responses emphasised the affective side of ownership and persuasion, others framed the issues more as teachers needing to understand the principles to experience success and make it work for them.

- *'I think it is more a case of the pressures on teachers and trying to convince them that by using these principles and techniques you actually get further on. So, it is not resistance, but it is a kind of case not being made just yet. Needing to go through a cycle of seeing it work. So, the research says it works, but it is teachers having to experience that for themselves' (Interviewee 1).*
- *'We still have staff turnover each year and one of the challenges is staff buy-in. It can seem a lot if you have not got any experience of it. Staff who are experienced are joining the school, they have been doing things their way for however long. The challenge is getting them to buy in. Once staff realise and they give it a go, that's when they'll see the progress students make … I guess it is about teacher buy-in and making it clear that this has an impact and it does benefit pupils more than other strategies … Fortunately, we have a young enthusiastic team at our school, people buy in to it quite easily. I don't know if other schools have it that easy. Also the SLT, that is a challenge, you may have a new headteacher ripping up what you have been doing the last five years. It is making it very clear that it is effective' (Interviewee 11).*
- *'Rather than feeling that things are imposed on you, teachers need to learn how children learn, how the brain develops, and how this will enhance their learning. The more we think this way, the better the children will learn. Teachers need to see that it is important, not extra work. As a profession we tend to think about it as extra work; it is not, it is thinking different about what we are doing' (Interviewee 7).*

There were also respondents discussing buy-in and understanding from the students:

- *'In my experience, I find students in exam classes don't like to experience the thinking or in other ways experience the "desirable difficulty" that comes with the mentioned strategies. Students felt worried when I changed my approach on teaching a topic where they were challenged every lesson. They preferred taking notes. My main worry as a practitioner is … how applicable are these strategies in the actual classroom?'*
- *'I think students should be taught best practice strategies and sold on the benefits, but do not need to know the science behind them.'*
- *'So I don't know if we are seeing it with the older students, there is a resistance because they are not used to learning like this, but maybe there is as time goes on and it is being used in primary and they are coming through and they are used to these strategies, maybe they will learn*

differently. I don't know. I think it is such a useful thing as a teacher to understand the science of learning and to know the process of memory and how that works. But from the students' point of view, if they are not used to using it, they are just seeing it as another thing they are being asked to do and they don't understand how useful it is and why it being done to them' (Interviewee 2).

Whole-school and departmental approaches

What is the role of school leadership in whole-school or departmental implemental approaches?

There were a number of responses relating to the value of a whole-school or departmental approach to implementation. Some of these related the approach to external and within-school professional learning:

- *'I think there should be a consistent whole-school approach to cognitive science strategies. This could provide more opportunities for teachers to collaborate and share best practice. I also think that trusted external training could help in reducing bias in some approaches.'*
- *'We have something in place about sharing ideas. I want to get to a point where staff meetings are about staff sharing ideas with each other. If someone goes on a course, they interpret that and bring their own things to it. So, consistency is about hearing the same thing' (Interviewee 1).*
- *'We have been developing cognitive science strategies as part of our improvement planning for the last few years. We have a strong senior lead teacher, who is also a NACE associate, and she is very proactive in sharing up to date strategies and processes. However, this is also something that the whole leadership team is supporting. To be most effective—and to have the greatest impact—this approach needs to be consistently reinforced and promoted by all if it is to become embedded across the school. I think we are getting there as a staff, but there is more to do.'*
- *'We are working on how we can apply this across all year groups, because if they haven't used them in earlier years, then we have a large catch-up when they get to Year 6. We need to build foundations, making sure we are constantly developing these strategies so it becomes second nature' (Interviewee 6).*
- *'A group of colleagues and myself started our journey into the use of cognitive science in teaching a number of years ago; we delivered whole-staff training on this (and shared this across the trust)—all new staff are introduced to the concepts. Many departments have now written these strategies into their schemes of work. It is a work in progress but it is beginning to become more embedded across the school.'*

Consistency and teacher autonomy

How much autonomy should teachers have in the selection and application of teaching strategies?

Above we touched on some of the challenges of promoting consistency while ensuring subject-specific aspects are attended to. While many respondents stressed consistency, others emphasised the value of teacher autonomy in the use of the strategies; others held a position between these.

On the value of consistency and the expectation that cognitive science informed strategies are used:

- *'A teacher selecting which strategies they'd like to use leaves it too open to the individual to not use these approaches in their classrooms. All initiatives we have implemented in our school have been consistently implemented across the school. We have done independent research and also*

had training together. We have agreed in staff meetings about what and how we are going to implement and this is monitored by the SLT.'

- *'I think it needs to go much more into early teacher education. I think personally that if something has been shown to work, it shouldn't really be a choice. If it is going to do the kids some good, it shouldn't be up to the school itself to decide' (Interviewee 8).*

Many made a distinction between foundations or principles and specific strategies:

- *'In regards to teachers having choice over which techniques to use, I believe all teachers should be aware of all strategies but should be given flexibility to choose the appropriate ones for subject and requirement.'*
- *'I do think teachers need some autonomy to choose strategies but the foundations should be evident in all lessons.'*
- *'I think it is important that cognitive science strategies are used consistently across school but how each of these strategies is delivered should be left to the strengths of each teacher. For example, retrieval practise should be visible in all classrooms, but this might be a "Quick Six" quiz in one classroom, a "Do Now" in another, a lollipop matching activity in a third etc. etc.'*
- *'Whole-school, consistent approaches are key to the success of these strategies. However, I also believe teachers should have a level of independence in how they implement the strategies within their classrooms.'*
- *'I believe that there should be consistency in strategy across the school in order for pupils to be able to identify patterns and learn how to learn. That being said, as with every aspect of pedagogy, all teachers should be able to adapt their practice as necessary.'*
- *'I believe that every teacher should have a choice in the implementation of these strategies. New teachers should be provided ample mentoring to ensure proper practice and senior teachers should be given appropriate training to stay up to date.'*
- *'I prefer to make independent choices about what I would like to implement in my classroom/year group. However, I have found that a school culture/expectation to use certain strategies gives teachers freedom to apply strategies and also encourages the use of these strategies. It also encourages investment in resources and training.'*
- *'I believe cog sci strategies are excellent but they need to be understood and implemented correctly. A blanket imposition on teachers who don't understand them is not going to work (this has been the problem in the past with things I have seen introduced, and currently see in some schools with KO's [knowledge organisers]). Likewise, "teaching" cog sci strategies to pupils needs thought—I explain them as and when I believe to be helpful, as we do something.'*

Continuing professional development and learning (CPDL)

What is the role of professional development and learning? Do teachers report that sufficient training is available?

Another frequent discussion point when we asked teachers about implementation of the strategies was continuing professional development and learning (CPDL), both in terms of the quality and availability of external training and in-school approaches to, and cultures of, professional learning.

Many emphasised the importance of training quality:

- *'How the strategies are introduced and used by teachers is crucial: I have seen so many things introduced badly by poor instructors or second-hand. For example, head does course and passes it*

on and that person doesn't explain the background or the "why" well enough ... so it becomes a "thing" teachers do rather them understanding why.'

- *'I think it's important that cognitive science is understood more broadly than just a checklist of strategies to implement. This helps to reduce the chance of so-called lethal mutations.'*
- *'I think it is just the quality of the teacher who is talking about the strategies and I think the quality of the time for staff to discuss these things in very busy times … they need to try it out. They need to then get some feedback on how that strategy works and how open you are to that sort of feedback depends on your personality, doesn't it? In order for these strategies to work it requires good continuous development of staff, but also staff willing to support each other and have those professional conversations with each other. So it is not senior leaders coming in, they are doing it amongst themselves. It's about having those sort of conversations' (Interviewee 1).*
- *'If you don't train teachers and students in the "why" and the "how" of how these things work then the chances of mutations are high!'*

Some discussed the availability of training to them. Several felt that training was not widely available and many mentioned twitter as a key source of information (but note that many of our questionnaire responses came from sharing the questionnaire online):[30]

- *'There has been little training available in the locality I work in on any of these areas.'*
- *'My knowledge of recent research and current ideas around evidence informed practice are largely drawn from my own reading, mostly via Twitter or Teacher Tapp. In addition, we have provided a member of staff the opportunity to become a maths mastery specialist and some of these strategies are filtering through into school in the context of mathematics in particular, for example, retrieval.'*
- *'At my school we don't really talk about cog sci and stuff and when I did my PGCE it was kind of new, four to five years ago. Most teachers seem to be engaging with through Twitter … interleaving cognitive load, etc. There are so many things I didn't know until I started engaging with it online. Started engaging with it in relation to curriculum planning. And what I have implemented from what I read' (Interviewee 12).*
- *'Engaging with wider reading, blogs, twitter etc. has revolutionised my teaching in the last 18 months. I'm really enjoying up-skilling myself and improving my practice.'*

New language for old practices?

In our questionnaire, we asked teachers to rate their level of agreement with the following statement: 'Cognitive science is a new way of talking about old teaching strategies.' Many 'somewhat agreed' with this statement, but there were a large range of responses, and many provided follow-up explanation in the open response section.

Some commented that cognitive science was showing new ways of doing things:

- *'I do think that data from cognitive science rescinds many assumptions we have about how children learn and so it's really important that we start to question our practices.'*

Many thought that the strategies were more longstanding:

[30] Indeed, one teacher we interviewed thought that the dissemination of teaching approaches on social media carried with it dangers: 'I think social media is a dangerous thing, because it's, if I look at my social media streams, it appears the world is getting very much research informed. Because of the way the algorithms work, you are drawn to the people who share similar things' (Interviewee 10).

- *'I am an "old fogey" but after training realised that I was doing all of these strategies but just did not name them as such.'*
- *'I think it is important to acknowledge that some teachers will be using cognitive science strategies without necessarily labelling them as such. There is good practice evident in many classrooms which relates directly to the principles of cognitive science but may fall under another name or not even labelled at all as they are just something that teacher does ... I was taught dual coding in my training year in 1990 but nobody ever called it dual coding.'*
- *'In regards to "new way of talking about old strategies" I agree with this as some of the techniques employed have been done by skilled teachers without knowing that is what they were doing explicitly.'*
- *'I have not heard of those specific terms before, but when training to be a teacher, learnt about those teaching strategies. I am aware that they are successful to children's learning and they are always picked up on as good practice.'*
- *'Much of this cognitive science is long established good primary practice; we call it different things but it delivers the same outcomes.'*
- *'I believe our holistic curriculum encompasses a lot of cognitive theory in its delivery. In some ways, the science is catching up to confirm the classroom practice.'*

Several discussed the continuity of old to new practice as a positive, with many placing specific value in the value of the language and ideas of cognitive science, even where strategies were more longstanding:

- *'It has been interesting to see how names have been given to strategies that older teachers have found intuitive. Having evidence to explain why those strategies work has been incredibly useful when discussing with students.'*
- *'Cognitive science is helping them have some framework for what they are actually doing. That's what I am saying, [it's] old-fashioned in a sense—things that good teachers did naturally in the past, like consolidation, which kind of went out of fashion for a few years, rather than new learning … [it's the] real understanding that children need time to actually consolidate and practice what they are learning' (Interviewee 1).*
- *'I think that the cognitive science gives new names to existing teaching practices. So, for example, retrieval practise, I think, went on all the time, but perhaps it gave me a better understanding of what I was doing ... often with teachers who have been in their profession for quite a long time but there's… I've always heard them saying there's a little bit of a kind of stereotype with them going "everything just comes round in circles, everything just comes back again" and I think that I've been in teaching long enough to see that happen and something gets recycled and you, like, you were doing that years ago. But I do think that the cognitive science… that these things are new, they are new and as much as they relabel existing good practice I do think that there's something to be learned' (Interviewee 4).*

One respondent felt that whether or not cognitive science matches up with older teaching strategies, it has little bearing on its value:

- *'Cognitive science isn't a new way to speak about old teaching methods; it is evidence-based research into what does and doesn't work, and why. If the findings match up with what teachers already do or did then that is a by-product, due to longevity of teaching as a profession.'*

Final thoughts

We close this section by reporting some general comments from the questionnaire followed by two extended reflections on cognitive science from our interviews. In these quotations, teachers are commenting on and considering the current status of applied cognitive science, its professional value, its dangers, and its potential.

Two positive comments:

- *'I find it [cognitive science] quite exciting because there's so many things that I'm just discovering and using and it also makes me feel like I've got more control ... It's something that has made my life easier ... cognitive science is absolutely your friend' (Interviewee 3).*
- *'Cognitive science has revolutionised my teaching. Some strategies can be put into place straight away, however others, I think, take more time to be implemented and embedded.'*

Two tentative comments:

- *'The only reservation I have is when it is treated as the holy grail. Without pupil motivation, group belonging, strong relationships, it can't deliver. Some of the strategies can be weakened by not being discussed or modelled in the context of real classrooms with real kids and real-world challenges. That doesn't invalidate what they offer, but can neuter their effectiveness.'*
- *'I think cognitive science is in the early stages and runs the risk sometimes of becoming a "buzz"—for example, someone will write a research paper into the impact of displays on learning and before you know it schools are ripping down displays without really understanding the science behind it.'*

And two negative comments:

- *'It has not made one iota of difference. Good teaching is good teaching, regardless.'*
- *'I think the research has been taken out of context and wildly exaggerated. There are many who fail to consider the research in any depth or nuance—particularly the limitations.'*

We close this review of perspectives from the practice review with two extended quotation from two teachers participating in our interviews: First, an extended quotation from interviewee 13 on the future of applied cognitive science:

> *'I'd like see it as something that is taught to teachers in a very intentional way, that what the thing I'm probably most interested in out of everything is teacher education. What we find is that in our school, and the reason we can do things as a whole school, is because we spend four or five hours a week training our teachers, and teachers working together on things, and collaborating in professional learning communities. And that, for me, is the root of it. I think there's a lot about at the moment—there's a lot of talk about cognitive sciences, a lot of talk about retrieval practice and stuff like that. You see it everywhere on Twitter and various things. But … teachers and school leaders have to have… they have to have access to the stuff that's … that's accurate and reliable, as opposed to anecdotally what someone thinks … And you get a kind of echo chamber going on, where you get lots of people agreeing with each other about the things that are right, and following certain types—certain people—as opposed to the research … I think that teachers need help with it. I also think that our understanding in teaching of cognitive science is probably quite narrow at the moment, I would imagine there's a whole lot more to it than we're aware of. But the thing that teachers go after is the retrieval practice and things like that, because that's something tangible that they*

can get hold of. And that that actually might not be the highest leverage thing. It might be that there's something else out in the ether, it's just that nobody knows about it.'

(Interviewee 13)

Finally, an extended quotation from Interviewee 9 on the future of applied cognitive science:

'In my school it is quite a big thing, just because we are all convinced that if we can make our lessons better, students will be able to learn more and in a more efficient way. It is going to make our lives easier because every lesson we teach will be more effective. I think in our school it is something that we will pursue and make more of. I think at the start of it, teachers who have been around a bit, can be a bit like, "Oh here we go again, it is another trend, this year it is all about cognitive science." I do think this is different, because if these strategies are what makes you learn… You know there was stuff in the past about learning styles, and the next year it is something else. Theoretically, if the learning scientist theories are true, they shouldn't really go anywhere, should they? We would be able to develop them. The way your brain works isn't going to change anytime soon.'

(Interviewee 9)

Part C: Conclusions

C1. Summary of results by area

Spaced practice

Brief definition[31]

Spaced practice applies the principle that material is more easily learned when separated by an inter-study interval (ISI). ISIs can be very brief (seconds or minutes) or very long (weeks or months). Spaced practice is also referred to as 'spaced learning', 'distributed practice', 'distributed learning' and 'the spacing effect'. Spaced practice is often contrasted with massed (or clustered) practice whereby material is practiced in a single session or close succession.

Spacing can be applied to many aspects of teaching and learning, including the spacing instruction or delivery (such as the information provided on a particular topic), practice (such as completing worksheets), or assessment (for example, the frequency of quizzes or formative tests). Spaced learning is one of several cognitive science informed strategies labelled as a 'desirable difficulty'; learning may be more challenging on a short-term basis but long-term retention is enhanced as a result (Greving and Richter, 2019).

Summary of results

We reviewed 27 studies focused on spaced practice. We identified two strategies for which we potentially had sufficient evidence to assess effectiveness. Our results for these are summarised in Table B1.8.

Table B1.8: Spaced learning—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
'Standard' spaces, across lessons or days	Eighteen, of which three were graded as high priority.[1]	The overall evidence suggests that spaced practice has a small but positive effect on learning compared with massed practice.	There was a good age range (6 to 17) represented. There were a range of subjects, including literacy, maths, science, and PE—although this was limited for larger and high priority studies (to maths, science, and critical thinking).	Low (++)
'Short' spaces, within lessons	Two, of which both were graded as high priority;[1] one a meta-analysis of six small-scale trials.	The evidence suggests a positive effect on learning compared with massed practice and that it might be a way of learning content in a highly time-efficient manner.	Outcomes were science, geography, and history, with a good range of ages although, there are too few studies here to reach a judgement about applicability.	Very Low (+)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

[31] For this summary section only a brief definition is provided. Full discussion of the definition is provided in the corresponding results sections.

Spaced practice—conclusions about strategies in this area

Our headline conclusions in this area are:

- Spaced practice is *potentially* highly relevant across the U.K. education system, for all learners and subjects. The spacing of learning is a fundamental aspect of curriculum and lesson design. Longer spaces affect curriculum design, within and across school years; shorter spacing is highly relevant to lesson planning and pedagogy; standard spacing is potentially relevant to both.
- The results suggest a small but positive effect for spaced practice (d = 0.2 based on the highest precision study and other results in the 0.1-0.2 range).
- The high priority and largest studies represent a more limited range of studies (maths, science, critical thinking), with a suggestion that effects in other areas are lower and/or less consistent.
- There was more evidence for 'standard' spacing (that is, across lessons and days) than within-lesson spacing. There was indicative evidence for the latter, and it follows from the theory, but there is too little evidence to reach a firm judgement.

The implications of the evidence presented above is that spacing is a plausible strategy for promoting additional learning. Still, there is large variation in the practice, and the evidence-base is currently relatively limited, even for assessing whether spacing is, in general, effective. More robust research on the overall effect of spacing and moderating factors is needed before firmer conclusions, and more confident recommendations can be made about whether to space learning and how to do it effectively.

While these results do not suggest a large impact of spacing, it was one strategy area that may be possible to implement at scale through building spacing into units/schemes of work at the planning stage. However, we also note that spacing may create further demands on an already crowded curriculum. This issue is one we explore further in our discussions and questions section for spaced practice.

Interleaving

Brief definition

Interleaving consists of sequencing learning tasks so that similar items are interspersed with slightly (but not completely) different types of items rather than being presented consecutively (Rohrer et al., 2019). When learning tasks are interleaved, they are inevitably also spaced. This can make spaced practice and interleaving hard to distinguish (Agarwal and Bain, 2019), practically and conceptually. In spaced practice, spaces are usually filled with unrelated activities or the learning of unrelated topics.

Summary of results

We reviewed 12 studies focused on interleaving We grouped these into a general interleaving area for which we potentially had sufficient evidence to assess effectiveness. Our results for these are summarised in Table B2.4.

Table B2.4: Interleaving—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Interleaving	Twelve, of which six were graded as high priority.[1]	For specific applications of interleaving in maths (relating to tasks involving strategy selection) the overall evidence suggests moderate to large effect sizes.	The studies spanned a range of ages from 8 to 14 and the vast majority were in maths. Generalisations beyond these ages and maths is not possible based on these results.	Low (++)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Interleaving—conclusions about strategies in this area

Our headline conclusions in this area are:

- The most notable aspect of this evidence-base is that 11 out of 12 studies were in maths. Moreover, even within these, the focus was on interleaving mathematic tasks that required learners to select a solution strategy before implementing it. This was true for five of the six high priority studies. The other study (Nemeth et al., 2019), while it did not test overall performance differences, found that interleaving improved the flexibility and suitability of students' strategies.
- The evidence supports the overall theory of how interleaved *maths* tasks may promote learning: with variation and the need to actively select strategies, students become more familiar with the differences between strategies, more able to discern between them, more confident in carrying them out and more discerning and flexible at selecting them.
- For this specific application of interleaving, the overall evidence suggests moderate to large effect sizes.

A question not addressed in this data, to which we return in the discussion and questions section is whether interleaving is likely to have value across other subjects and applications. We review literature recommending the application of interleaving across the curriculum and discuss the theory and practice in this area. Finally, we return to the opening comments in this section about the connection between interleaving and spaced practice, in terms of timing, and comparison (a strategy in the Working with Schemas section). Assessing evidence in these three sections side-by-side is valuable for increasing understanding of this area.

Retrieval practice

Brief definition

Retrieval practice 'refers to the act of recalling learned information from memory (with no or little support)' (Jones, 2019). Principles of learning from cognitive science suggest that learners actively generating responses from memory and quickly receiving feedback will be an effective learning approach. A common way of achieving this in a classroom is through low-stakes quizzes, questions, and tests.

Summary of results

We reviewed 21 studies focused on retrieval practice, of which 16 tested retrieval practice against restudy or re-presentation of material. Our results for these are summarised in Table B3.5, below:

Table B3.5: Retrieval practice—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Retrieval practice (compared to restudy or re-presentation)	Twenty-one, of which three were graded as high priority.[1]	The overall evidence suggests that retrieval practice is an effective learning approach per se (i.e., against a no-treatment condition). Against restudy or representation of material, we judge there to be a positive effect overall, indicating moderate effect sizes.	A good range of learning areas was examined within the studies. The learning outcomes tended to be a factual recall or vocabulary learning, although there were a small number of examples of learning with higher 'element interactivity', where elements needed to be connected as well as recalled.	Low (++)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Retrieval practice and the testing effect—conclusions about strategies in this area

Our headline conclusions in this area are:

- Retrieval practice is highly relevant across the U.K. education system, for all learners and subjects.
- The findings in this area are mostly positive, suggesting moderate effect sizes, but there were an appreciable number of neutral or negative results.
- There was a good range of subjects and ages represented in the results. This suggests that the principle might have wide applicability across curriculum areas.
- One issue in this area has been the low ecological validity of the studies. The vast majority were designed and delivered by researchers, often in schools but outside of the classroom. Moreover, many interventions have been wholly scripted with a standardised procedure. This raises questions about whether 'real' teachers are likely to achieve the same results in more realistic conditions. The results from Churches et al. (2020), one of our high priority studies suggest, so but a firm conclusion is not possible based on the limited evidence we have in this area.
- We have focused specifically on studies testing retrieval practice against restudy or representation of material. However, we would note that re-study or re-presentation is likely to also result in learning. Using this as a comparator is perhaps a demanding test of the strategy. If we look at all 21 studies, this would add (with five studies with positive results) further weight of evidence to the conclusion that retrieval practice is effective.

Finally, we note that a recent systematic and meta-analytic review of testing, looking across all age ranges and a wide range of contexts also proves additional weight to these conclusions (Yang, Luo, Vadillo, Yu, and Shanks, 2021). Yang et al., (2021) was published during the final stages of the write up of the present review. This study estimated a medium overall effect of testing (quizzing) (g = 0.50, CI: 0.44, 0.56) and also provides support for the use of different test formats and corrective feedback. As well as noting the main result here, we reference several of the findings from Yang et al. (2021) below where they provide further evidence in response to several questions we pose.

Managing cognitive load

Brief definition

Detailed or complex presentation of information or problem spaces can easily overwhelm the working memory. Thus, managing cognitive load is not a task of *minimising* cognitive load but rather *optimising* it. For optimal learning, educators should—so the theory goes—look to minimise extraneous load

while maximising intrinsic (or germane) load while not exceeding the working memory capacity. Evidence also shows that there is substantial variation in working memory capacity between individuals (Cowan, 2016; von Bastian and Oberaurer, 2014); effective cognitive load management will therefore require consideration of specific pupils and pupil groups and is likely to pose a greater challenge when teaching a full (and especially a mixed-ability) class.

Exceeding working memory capacity is a particular issue in relation to 'problem solving' in education, where learners are typically presented with a large amount of information in quantity or complexity and asked to successful identify target information or follow (or sometimes discover) a series of problem-solving steps (Sweller, 1988). Students, especially those with limited prior knowledge, often struggle to navigate through this problem space such that working memory is overwhelmed and learning is impaired. In response to the limitation of working memory, there are educational strategies that seek to reduce or optimise the load on working memory.

Summary of results

We reviewed 91 studies focused on the management of cognitive load. We identified three strategies for which we potentially had sufficient evidence to assess effectiveness. Our results for these are summarised in Table B4.9, below:

Table B4.9: Managing cognitive load—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Worked Examples	Twenty-two, of which four were graded as high priority.[1]	Small to moderate positive effect of using worked examples compared to conventional problem-solving techniques	Results were entirely concentrated in maths and science and secondary-age students (11–18 years old).	Moderate (+++)
Scaffolds, Guidance and Schema-Based Instruction	Sixteen, of which two were graded as high priority.[1]	Well-targeted scaffolds, guidance or schema-based supports are an effective approach to support students to solve problem or learn from complex tasks.	There was a good range of students from age 8 to 16. Most studies were either maths, reading comprehension, or science, with a roughly three-way split between these.	Moderate (+++)
Collaborative problem solving with worked examples or scaffolds	Nine, of which one was graded as high priority.[1]	The evidence is supportive of the theory that collaborative learning will lower cognitive load and support learning during problem solving or complex tasks; although there were some negative results and complexity.	Student ages ranged from 8 to 16. Most of the studies were focused on mathematics learning (six). There were 2 science (biology) and 1 ICT.	Low (++)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Managing cognitive load—conclusions about strategies in this area

Our headline conclusions in this area are:

- Cognitive load has high potential relevance across the U.K. education system and for all learners and subjects.
- Overall, the evidence is promising and indicates the value and importance of teachers seeking to optimise learners' cognitive load.
- There are numerous studies showing appreciable positive effects for strategies to manage cognitive load within the evidence we have. There are also appreciable numbers of neutral and

negative results, suggesting complexity in the principles and challenges of making it work in practice.

- Much of the evidence we have is highly concentrated in specific age ranges and subject areas. Tests of worked examples have almost exclusively focused on secondary maths and science.
- Considering worked examples and other forms of scaffolding (for example, support and guidance for complex learning or problem-solving spaces) together suggests wider subject and age applicability (age 7 to 16) of the principle and provides greater confidence in the overall result. However, we note that this confidence is in the value of optimising cognitive load *per se,* rather than a specific strategy for doing so or for specific learner needs.
- Ecological validity was low for many studies, limiting our ability to generalise the findings to real educational settings confidently.

Worked examples

The evidence was largely in line with the overall theory but suggests that as learners develop knowledge, only partial supports are required. It can be challenging to consistently identify best practices. For novice learners, however, the evidence is clearer and supports the use of worked examples to manage cognitive load and support learning.

There are many studies in this area, but there are limitations in their robustness (vis-à-vis internal validity) and ecological validity. The other limitation with *worked examples* is that all 22 studies we reviewed were studies of mathematics (17) or science (5), and the majority of studies were for secondary-age students (20/22). Thus, while the results support the use of worked examples in preference to unguided problem solving in secondary maths and science, we must stress that the limitations in the present evidence-base prevent judgements of effectiveness beyond these subjects.

We also examined *incomplete and incorrect worked examples* within the overall worked examples section. The overall theory suggests that as learners start to develop knowledge in an area, incomplete and erroneous working examples can increase (desirable) difficulty and enhance learning. Our results are, again, broadly supportive of this principle (in secondary maths and science) but the results were less consistent than for worked examples as a whole. There appear to be issues matching learners with the right level of support. Moreover, many of these studies did not provide a breakdown of students' abilities and so we cannot make a confident judgement about whether student ability or their developing knowledge in the problem area is a key moderator of the effect as hypothesised for these studies.

At the outset of the managing cognitive load results section, we describe how the theory relates to the *optimisation rather than minimisation or maximisation of working memory load*. Incomplete or incorrect worked examples will tend to lessen learners' cognitive load compared with unguided problem solving but produce a higher load when compared with complete worked examples. According to theory, whether this is optimal significantly depends on pupil prior knowledge in the problem area. The mixed results in the incorrect and incomplete worked examples section can therefore be interpreted as being in line with the overall theory. Still, the evidence is limited and suggests that it is difficult to make work in practice.

Scaffolds, guidance, or schema-based supports

The evidence suggests that *scaffolds, guidance, or schema-based supports* effectively support students to solve problems or learn from complex tasks. A wider range of pupil ages and subjects were represented in this data giving us greater confidence that the strategy is more widely applicable. However, the downside of this diversity was high heterogeneity in the learning aims, subjects,

procedures, and assessments within this group of studies. The grouping of these studies was on a conceptual rather than practical basis. The practices were very different but we judged (prior to analysis) that all studies focused on learning complex material with supports designed to lower cognitive load (but not specifically focused on the provision of worked examples, as per the previous strategy). The main groups within this are (a) providing targeted explanations to support learning, (b) providing schemas and structures to support students to manage tasks, and (c) providing supports that manage information during the activity. Our overall 'moderate' confidence in our judgement that this is an effective general strategy comes with the caveat that we have specified the strategy at such a general level that it encompasses a huge range of practical strategies.

The other consideration is *how* the various supports used in this group of studies are conceptually and practically similar to those examined for worked examples. There were certainly many surface similarities, and it might be argued that some of the scaffolds we looked at in this section were the equivalent of worked examples—in particular, incomplete worked examples—but for a wider range of subjects. Subjects outside maths and science often have learning content that does not lend itself to specific and distinct (or algorithmic) problem-solving processes, for example, and so scaffolds, guidance, and schema-based support might be needed to manage cognitive load effectively. If this parallel is reasonable, we might look at the evidence in both areas collectively (note that both had an overall positive result with moderate confidence). Our categorisation of these studies was conducted before the analysis and separated out these two strategies. Future work, with greater attention to the specifics of strategies used, may wish to consider these collectively within a more granular taxonomy of the strategies and their contexts.

Collaborative problem solving

Finally, in relation to *problem solving*, our results suggest (a) positive effects of collaboration during traditional problem solving, (b) that complex or erroneous worked examples are best for individual learners, and (c) that worked examples with incomplete knowledge are superior for groups. However, we note that this summary is based on a small evidence-base with particular limitations in relation to pupil age and subject. Our confidence in this finding is low. Our judgement is that the complexities of task demands and the dynamics of group learning make clear principles about effective strategies more challenging. There is good evidence here that working collaboratively can lower cognitive load. Whether this optimises it for all learners, and the principles of how to do so, is a question that goes beyond the limited evidence we have and is a question we return to in the discussion and questions section.

Working with schemas

Brief definition

Knowledge is organised in the mind into connected frameworks of information known as schemas. In this area, we reviewed all studies in our database that focus on representing and developing schemas. There are a group of learning theories and strategies that seek to elicit or represent schemas as a way of presenting connected ideas, identifying a learner's pre-existing knowledge, and developing this knowledge, often through working with schematic representations of it as scaffolds to manage cognitive load and emphasise pertinent information. Working with schemas often involves developing ideas through processes of organisation, comparison, or elaboration.

Summary of results

We reviewed 25 studies focused on the organisation and comparison of information. We identified two strategies for which we potentially had sufficient evidence to assess effectiveness. Our results for these are summarised in Table B5.6.

Table B5.6: Working with schemas—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Concept/ knowledge mapping and organisation	Fifteen, of which three were graded as high priority.[1]	The evidence was mixed. Overall, it was positive, (12/17 results) but the negative studies suggest caution is needed.	Most studies in this group were for students of late primary to early secondary age (12 of 15 studies for age 8 to 14). Most studies were focused on the organisation and study of text using concept maps.	Very low (+)
Schema/ concept comparison and cognitive conflict	Ten, of which one was graded as high priority[1]	The overall results suggest promise for KS3 science and maths, however, the evidence-base is small and provides mixed results.	All studies were for maths or science, with the vast majority of students in the 11–14 age range.	Very low (+)

[1]High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2]based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Working with schemas—conclusions about strategies in this area

Our headline conclusions in this area are:

- Concept mapping and, more generally, the comparison of strategies and concepts have wide relevance for education for all learners and subjects working in areas with complex and connected information.
- The evidence presented, and its limitations, means that this area suggests both promise and pitfalls and raises as many questions as answers. Each area provided numerous discussion points that we considered further in the discussion and questions section.
- For concept mapping and the organisation of knowledge, there appear to be several variables at play, notably, the organisation of knowledge, the engagement with organised knowledge, and the extent to which students have generated or organised the representation (for example, a concept map). Our tentative conclusion is that concept mapping and organising knowledge can be effective approaches but that student-generated approaches risk excessive cognitive load or inefficiency (with time spent organising rather than active engagement with material) and benefit from retrieval or self-explanation scaffolds.
- Similarly, for cognitive conflict and comparison, the neutral and negative results all provide examples of studies where the level of support, engagement, and generation appear to have been pitched incorrectly given the learners' prior knowledge.

Cognitive theory of multimedia learning (dual coding)

Brief definition

We originally planned to focus this section on dual coding theory. Given the breadth of study strategies and questions, however, we defined this section according to a slightly broader theory: the cognitive theory of multimedia learning (CTML). CTML (Mayer, 2005) builds on ideas of dual coding, cognitive load, and generative learning with three assumptions that underpin the theory: (a) that working

memory has two separate channels for information (dual coding theory), (b) that each channel has a finite capacity, and (c) learning is an active process of working with this information.

Summary of results

We reviewed 55 studies focused on the presentation of information in multiple modes. We identified three strategies for which we potentially had sufficient evidence to assess effectiveness. We summarised these results in Table B6.8.

Table B6.8: Strategies related to the cognitive theory of multimedia learning—summary results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Visual representation and illustration	Thirty-four, of which three were graded as high priority.[1]	The evidence suggests that visual aids are most helpful during learning but frequently have no effect, and can sometimes be harmful.	The age range of students was 7 to 18, with a good spread within this range. Note that there is therefore no evidence for children younger than this. Studies represented a range of subjects. Although over two-thirds were of maths and science.	Low (++)
Diagrams	Fourteen, of which three were graded as high priority.[1]	The evidence suggests that diagrams for secondary maths and science learning are mostly helpful, but frequently have no effect and are often harmful.	Twelve out of the 14 studies in this area were for students between the age of 12 and 16. Also, 13 out of 14 were in maths and/or science. In this sense, the sample is relatively narrow.	Low (++)
Spatial, visualisation and simulation approaches	Seven, of which two were graded as high priority.[1]	Overall, this area shows some promise, but the evidence is insufficient to judge the effectiveness of strategies in this area, either for primary maths or more widely.	All studies in this area with one exception were focused on the effect of spatial visualisation in maths, including geometry and number. The student population for these studies spanned ages 4 to 12 and therefore represents the primary, but not the secondary age range.	Very Low (+)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Dual coding and multimedia learning—conclusions about strategies in this area

Evidence about how teachers use visual information and combine modes of information has high potential relevance across the U.K. education system for all learners and subjects. The simple changes, for example, of adding or taking away images from instructional materials, replacing written text on slides with just images, or providing diagrams, could have far-reaching implications.

In terms of the evidence we have here, however, firm conclusions have been challenging. For example, for visual aids and diagrams, when we crudely compare conditions with and without these, there are mixed results.

However, a slightly more nuanced interpretation of the theory would hold that the impact of images would depend on their decorative or informational content, their role and centrality within the learning, the format and content of other modes of information and how complementary these were, how the image was engaged with (including student generation), the student prior knowledge, the overall cognitive load, and more.

The evidence is not sufficient for us to make these distinctions and reach robust judgements on the effect sizes for subgroups and their impact on different learning outcomes and populations.

We noted also that we located studies that compared the effect of images with audio or animations on learning. The former of these is arguably more relevant to dual coding theory than some of the studies that combine written text (visual) with images (visual), although, as we discuss, drawing on the cognitive theory of multimedia learning. Elsewhere, the simple equation of information presentation types with working memory processing of this information is complex.

Overall, it has been disappointing that in an area of evidence where we originally identified 122 studies that there are so few clear and robust tests of the theoretical principles in applied settings.

Embodied learning

Brief definition

In the scoping and protocol development for this review, we searched for concepts and strategies from cognitive psychology and neuroscience that may have implications for classroom practice. This identified a wide range of strategies informed by cognitive science beyond those more commonly represented in policy and practice sources we reviewed**.** One of these was 'embodied' learning, in which studies examined questions such as how enacting concepts, gesture, tracing, and actions could support learning.

Summary of results

We reviewed 14 studies focused on embodied learning which we grouped into a general strategy group for review. Our results for these are summarised in Table B7.4.

Table B7.4: Embodied learning—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Embodied learning	Fourteen, of which one was graded as high priority.[1]	Evidence in this area is consistently positive, with a range of small to large effects estimated. The evidence was quite limited, but suggests promise for gesture, tracing, and physical activity and play.	Our sample spanned the primary age range and into the early secondary range (age 5 to 14). A range of subject areas were represented providing a tentative suggestion of more general applicability across subjects.	Low (++)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Embodied learning—conclusions about strategies in this area

Our headline conclusions in this area are:

- Eleven studies in this area reported causal evidence, all of which found embodied learning more effective than a control group.
- Embodied approaches to learning show promise for primary and early-secondary education.
- Many studies found moderate to high effect sizes as well as some smaller positive results. The single study rated as high relevance and quality in this area was Margolin et al. (2020) who found an effect size of d = 0.37 in a study of embodied and play-based learning in physics compared to the normal physics curriculum.
- The evidence-based in this area was, however, limited. In particular, there were specific issues with ecological validity for studies in this group, with most interventions being researcher-designed and delivered.

- This, and the potential limitations stemming from lack of targeted searches in this area (studies were located through more general search terms), lead us to rate our confidence in these conclusions as low.

Mixed strategy programmes

Brief definition

A relatively small group of studies evaluated programmes where two or more of our focus cognitive science strategies were combined. Where studies separated out strategies through, for example, multi-arm trials or multiple experiments, we were able to include the studies in our analysis of specific areas. Where only combined results were reported for the effect of multiple cognitive science concepts, we have included this here as a mixed strategy programme.

Summary of results

Eight publications were reporting tests of mixed strategy programmes; seven of these were substantive studies. Of these, four were graded as high relevance and quality. Note that two studies, Schunn et al. (2018) and Desimone and Hill (2017), report the effectiveness and implementation respectively of the same overall programme. Our results for these are summarised in Table B8.4.

Table B8.4: Mixed-strategy programmes—summary of results

Strategy	No. of studies	Finding	Applicability of evidence	Confidence level[2]
Mixed Strategy Programmes	Seven, of which five were graded as high priority.[1]	Overall, the evidence provides a mixture of mixed/neutral to small-moderate positive results for programmes of mixed cognitive science strategies. There were suggestions in several studies of issues with implementation. We judge the effect to be highly dependent on programme quality and implementation.	The studies were all focused on maths and science, in the U.K. or U.S. and students aged 11–14.	Low (++)

[1] High priority papers potentially provided strong evidence and were selected for in-depth analysis.

[2] Based on an adapted GRADE approach, drawing on a risk of bias assessment for high priority studies.

Mixed-strategy programmes—conclusions about strategies in this area

Our headline conclusions in this area are:

- Our analysis concerned programmes testing two or more cognitive science principles combined. These programmes typically revolved around curriculum (re)design accompanied by professional development in the cognitive science principles and (to greater and lesser levels of success) implementing the materials.
- Overall, the evidence provides either positive or mixed/neutral results for programmes of mixed cognitive science strategies.
- The evidence presented above shows that, at present, there are few or no large-scale programmes that have been trialled and found to be effective. Moreover, those that have been trialled—of which there are only a small handful of ecologically-valid, rigorous examples—some have yielded disappointing results.
- Of the five strongest studies, three had neutral or mixed effects and two had positive effects. One positive result was based on a curriculum redesign delivered at scale through professional development. The other, positive result was based on cognitive science principles built into a

computer revision programme and, while promising, had lower ecological validity than other studies.
- There were suggestions in several studies of issues with implementation.
- Our confidence in the effect estimate is low.

Mixed-strategy programmes have high potential relevance across the U.K. education system, for all learners and subjects. As (or if) cognitive science strategies are held to be individually effective, combining two or more strategies in a single intervention might be expected to increase the overall impact as multiple, individually-effective strategies combine for collective and additive benefits. Moreover, as discussed further below, mixed strategy programmes are likely to be an important vehicle for applying and scaling cognitive science informed practices.

Overall, the evidence on mixed strategy programmes presented in this section yields disappointing results. As we discuss at length in the discussion and questions section, our reading of this area is that the effectiveness of these programmes has been determined as much by their programme design and organisation as their underlying teaching and learning principles. Small or null results may stem from the operational issues as much from the (in)effectiveness of the underlying strategies; the evidence we have is not sufficient to support either explanation. There are known issues with implementation in these programmes (as we discuss). However, these did not apply to all programmes, and the relationship between implementation success (in terms of fidelity and dosage) and outcomes is not consistent across the group of studies.

These programmes are likely to draw on principles and practices relating to school improvement, curriculum development, and effective professional development; the success of these at this organisational and policy level is likely to be important for the success of any intervention at scale. The successful design of school improvement and professional development programmes lies beyond the focus of the present review; nonetheless, we conclude that these will be important considerations if and when cognitive science programmes are tested or delivered at scale.

Agreements and disagreements with other reviews

We close this results section by considering how this review relates to the most relevant cognitive science review identified in our scoping and planning work, Howard-Jones et al., 2014 (see Section A1 for a summary). Howard-Jones and colleagues identify 18 topics within their review and, for each, assess the strength of evidence for educational effectiveness (high/medium/low) and distance to application (near/moderate/distant). A summary of the results (see Howard-Jones et al., 2014, p.8–12) is provided in Table C3.1.

Table C3.1: Summary of findings from Howard-Jones et al., (2014)

Topic	Strength of evidence for educational effectiveness	Distance to application
1. Mathematics—non-symbolic and symbolic representation of number	Medium	Moderate
2. Mathematics—finger gnosis	Medium	Near
3. Mathematics—mental rotation skills	Low	Distant
4. Mathematics—maths anxiety	Medium	Near
5. Reading	Medium	Near
6. Exercise	Medium	Near
7. Sleep, nutrition, and hydration	Low	Near
8. Genetics	Medium	Distant
9. Embodied cognition	Medium	Moderate
10. 'Brain training' of executive function	Medium	Moderate
11. Spaced learning	High	Near
12. Interleaving	Medium	Moderate
13. Testing	High	Moderate
14. Learning games	Medium	Moderate
15. Creativity	Low	Moderate
16. Personalisation	Low	Moderate
17. Neurofeedback	Medium	Moderate
18. Transcranial electrical stimulation (TES)	Medium	Distant

Comparing this summary to our own results, we have the following observations:

- Our review has organised the topics differently. We have highlighted in green the areas that have directly aligned. Howard-Jones and colleagues ('the former review') situated many areas within maths and English (topics one to five) whereas we have not organised our areas to be subject specific. In practice, however, many of our results have been concentrated within particular subjects. Topics such as personalisation, creativity, and learning games tended to crosscut other areas. The former review also represents a wider range of cognitive science strategies within the main review. For areas such as genetics and transcranial electrical stimulation, while these were considered in our scoping review, the research and practice on their application to practice was judged to be too limited at this stage to include these in our focus review areas.
- In terms of alignment of results:
    - we also find evidence for non-symbolic and symbolic representation of number in our 'spatial learning' and scaffolds and guidance strategy reviews;
    - we discussed indicative evidence relating to anxiety but not specifically related to maths and without sufficient evidence to reach a firm judgement—there is evidence that anxiety is an important cognitive load factor (and we also touch on this in our discussion of retrieval practice); and
    - as in the former review, we provide evidence in support of embodied cognition and spaced learning and, to a slightly lesser extent, interleaving and the testing effect.

- As well as different categories and review boundaries, we have used a different review methodology and systematic analysis tools, so it is difficult to directly compare results. Nonetheless, we view our results to be broadly in line with, and to build on, those from the former review.

C2. Implications and overall findings

In section C1 we provided a summary of results for each of the eight review areas and discussed implications of specific results at the level of each cognitive science area and strategy. In this section we consider over-arching findings and implications; we provide a series of review-level headline findings each followed by short discussion.

Overall findings

1. *Cognitive science principles of learning can have a significant impact on rates of learning in the classroom. There is value in teachers having some working knowledge of cognitive science principles. They should also be made aware of the serious gaps and limitations in the applied evidence-base, the uncertainties about the applicability of specific principles across subjects and age ranges, and the challenges of implementation in practice.*

There is enough evidence in the applied cognitive science evidence-base to conclude that cognitive science strategies and principles are significant factors affecting rates of learning and its retention in many everyday classroom situations. In our findings by strategy, we were able—at least tentatively—to support most of the strategies reviewed; however, this support invariably came with caveats relating to uncertainties about applicability across subjects and pupil ages as well as potential issues with implementation. We also would stress that cognitive science principles are not the only important factors for learning; we have not sought to compare their impact against, or in connection to, other influences such as classroom relationships, social-emotional learning, feedback, and so on.

We believe that there is value in teachers being trained in cognitive science principles—undertaking professional development and learning in the area—and their application in the classroom. We judge the evidence-base to be sufficiently broad and longstanding to dismiss the view that cognitive science principles are a transitory 'fad'. Many of the strategies are likely to be already in use without explicitly being articulated in terms of cognitive science. There is, however, value in a 'common language' and in making and maintaining connections between the evolving (applied and basic) science and education practice. That said—as we discuss below—we hold that the applied evidence has serious weaknesses and recognise that the science continues to evolve.

2. *There are large disconnects between the evidence-base for basic cognitive science and applied cognitive science. Applied cognitive science is far more limited and provides a less positive, and more complex, picture than the basic science.*

The sources we consulted during our practice review present a confident (and sometimes strident) account of the theory, practice, and value of cognitive science in the classroom. Many accounts we have reviewed in the practitioner-facing 'popular' cognitive science literature are presented with high confidence: the theories are highly elaborated and positive and there are many examples of the prescription of specific strategies over others; the accounts have high resolution and extensive detail, are comprehensive, often with high coverage across pupil ages and subject areas.

In contrast, the applied evidence-base that we have systematically reviewed supports cognitive science strategies with moderate, low, and sometimes very low confidence. While this lack of confidence stems mostly from paucity of evidence, it also reflects numerous areas with mixed results

that do not provide definitive support for the core cognitive science principles; the applied evidence is low resolution, complex, and many conclusions have been at a quite general level (that is, providing support for a general principle but not a specific strategy for addressing it). Equally, many of the results were promising, with small and moderate positive estimates of typical impact (but with some negative results).

The contrast between the level of confidence and strength of the applied evidence and most practitioner-facing accounts of cognitive science was stark. The limitations in the applied evidence are strikingly clear in the database statistics for our systematic review. We located, through our searches, over 40,000 individual studies. After screening these using our priority criteria, our database stood at 499 studies. Of these, only 43 were rated as 'high priority' studies with high potential to shape our findings, ecological validity, and relevance. After a risk of bias assessment, only 17 of these were rated as having low risk of bias overall[32] and four of them were rated as having 'high' risk. These 43 high priority studies, of which 17 had low risk of bias, were spread across 14 strategies in eight review areas—in other words, thinly. There were no areas that had a large number of strong studies. When we assessed groups of studies for each strategy using an adaptation of the GRADE assessment approach, we rated our confidence in the result and effect size estimates for four strategies as 'very low', seven as 'low', two as 'moderate', and none as 'high'.

Comparing this incomplete and uncertain picture to current policy and practice, and the accounts of practice-facing guidance based on the basic science, the pertinent question is whether these are in proportion to the evidence-base? Overall, the applied evidence is promising and supportive, *but also uncertain and hesitant*. In our view, the implication of this is that caution, nuance, and reflection are needed rather than prescription, simplification, and the blanket imposition of the prevailing conceptions of best practice across subjects, age ranges, and contexts.

3. The applied literature has many gaps relating to subject areas and age groups.

As touched on above, a key feature of many of our results is that we have had to qualify our assessment of the strategy with its applicability in terms of age group and subject. In general, secondary mathematics and science were areas well-represented across the evidence-base. There were, in contrast, areas where primary-age children or other subject areas comprised the bulk of the evidence, as well as strategies for which a good range of ages and subjects was represented. Rather than these being minor details for the conclusions, these gaps have significant implications for the confidence, scale, and areas in which cognitive science can be implemented while remaining evidence-based (at least in terms of the applied evidence). These gaps raise interesting research questions and also practical questions about how the nature of different subject areas and their curricula might make the strategies more or less appropriate and effective. For example, in science and maths, the technical and often procedural nature of knowledge and the attention to addressing misconceptions might increase the value of interleaving and the use of worked examples.

We also note, in the section on Embodied Learning, that that particular area is not well represented in the practitioner-focused practice review literature. More generally, we note that emotional and social aspects of learning are not as prominent in the prevailing popular accounts of cognitive science. We found that the current evidence-base on cognitive science applied to the classroom is more focused on how individual learners process and remember information; social, emotional, and physical aspects to cognition and learning receive less attention. This does not appear justified by the

[32] These low ratings resulted even when disregarding risk of bias assessment criteria about pre-planning of analysis. With these criteria included, this number would be far lower.

basic cognitive science evidence and what we know about the brain and learning. Our survey of teachers, review of the underpinning science, and the applied evidence in areas such as embodied learning suggest that social, emotional, and physical aspects to cognition are also important considerations for research and practice.

4. *Applied research surfaces many theoretical and practical problems not encountered in controlled lab or pseudo-lab conditions.*

Following from the previous point, the major factor in the large discrepancy between our account and those from our practice review was our focus on the *applied* cognitive science. A key assessment criterion in our review was the 'ecological validity' of the studies. Put simply, we wanted to know whether cognitive science techniques work in real classrooms, across the curriculum, and for different pupil groups. We discuss the question of ecological validity, applied research, and challenges of implementation at length in the discussion to the Mixed Strategy Programmes area. There we frame this issue as follows:

> *The problem of translating cognitive science principles into teacher practice at scale was not the focus in many studies in previous review sections. Most problematic for inference about transfer and scalability are intervention 'set pieces' delivered by researchers or experts, or scripted lessons or computer programmes for independent study. From the perspective of assessing efficacy or experimentally isolating cognitive scientific principles, there is clearly huge value in these studies, but for our present purposes of assessing the implications of the evidence for teacher practice, it is important that we also consider the necessary steps to get from a 'proof of principle' to a strategy suitable for widespread implementation by teachers.*

While largely in line with the cognitive science principles, and providing positive results, the applied evidence has been markedly less positive than the basic science evidence and popular accounts of it. In many ways this is a serious concern as it suggests that widespread implementation—and certainly that based on over-confident accounts of 'what the research says'—is premature. The applied evidence we have reviewed has surfaced many important theoretical and practical complexities in the implementation and understanding of the cognitive science strategies. Moving from efficacy (strategies working in controlled contexts) to effectiveness (with evidence that they work in realistic and multiple settings) is valuable for (a) providing confidence in the value of widespread implementation and (b) in terms of identifying the cognitive scientific and wider classroom factors that teachers will need to consider when incorporating techniques into their practice. As we have discussed in the various strategy areas we have reviewed, the various cognitive science strategies have suggested different practical considerations and challenges, in nature and in quantity.

5. *The evidence-base is largely working at the level of principles rather than tests of specific classroom strategies. Principles do not determine strategies and do not determine specific approaches to implementation.*

One subtlety in the conclusions is that many strategies have been stated at the level of principles. The clearest example of this was in the managing cognitive load section where our GRADE analysis identified two effective strategies in which the evidence was rated with moderate confidence. Here, we concluded that the evidence supported the *principle of teachers managing cognitive load* but that we were less confident in a specific strategy for doing so.

To a large extent, needing to work at the level of principles has been necessitated by the grouping of studies into strategy areas for assessment. Within each group, there was often considerable variation in the specific teaching and learning strategies used, the learning outcomes, and the context. Much of our analysis has wrestled with whether studies are sufficiently similar to provide a test of a strategy or principle while grouping to prevent the details becoming too particular and the groups too small and disparate. The evidence supports the view that these details, the context, and particularly the subject, matters. The implication for teachers from these results is that the research is not at the stage (and may never be at the stage) where specific strategies or approaches to implementation can be prescribed. Even with robust principles and strong understanding of these, teachers will have a considerable amount of work ahead of them to make them work for a particular learning outcome in a particular context.

6. Principles tended to be clustered and interact in complex ways.

In general, we have evidence for very few simple principles ('do A rather than B'). Our conclusions about cognitive science principles have tended to be either (a) conditional—where we conclude that they are effective if/but/when/for… or (b) linked to other, sometimes complementary, sometimes countervailing, principles. This aspect of the evidence is particularly evident in our discussion sections where we discuss the numerous basic scientific and pedagogical principles that are thought to affect the success of the strategies. There are implications here for teachers and for researchers. For teachers, the implication is similar to the previous point: that teachers are likely to require significant expertise to apply multi-faceted and connected theories in the classroom. This also creates a challenge for research. As noted above, most of the research has focused on isolating and examining single, simple principles in semi-controlled settings. Some of the most interesting and informative research we reviewed brought several potentially competing principles together (using, for example, factorial designs) to attempt to tease out how these work as a group and can be combined in pedagogy in realistic settings. There is great value in basic and applied research seeking to move from the urge to isolate and demonstrate principles to better understanding multi-component theories and clusters of principles working in tandem.

7. There are important connections across the cognitive science areas and strategies.

In most areas, we have indicated the connections between the strategies and principles reviewed. In many cases there were considerable challenges in delineating areas across the literature. This was a challenge from the perspective of the implementation of this review, but we view it as having positive implications for teachers and future research, namely, that there is value in researchers further examining the connections and interrelationships between cognitive science principles and strategies. There appear to be particularly strong connections between three of the groups. First, the areas concerned with practice including spaced practice, interleaving, and retrieval practice: many teachers described these in the same breath and there may be value in both connecting and discriminating between these in a single account. The second apparent group concerns the presentation of information. This group connects the cognitive theory of multimedia learning, managing cognitive load, and the embodied learning areas: all provided principles for the provision of information within instruction and to support student work. Third, there are a group of studies that are concerned with the 'middle ground' between fully-guided and didactic instruction and the student's role in learning. This concern is centred on our Working with Schemas section and cross-cuts the more learner-led, learner-generated aspects of strategies in the managing cognitive load, cognitive theory of multimedia learning, and embodied learning sections.

At a more granular level, there were many links within and across areas, strategies, and these broad groups outlined here. Future research and practice would benefit from discerning, where connections are identified, whether mechanisms and principles are (a) common, (b) complementary, or (c) countervailing.

Concluding statement

Based on the findings of this systematic review of the evidence, we are convinced that basic cognitive science and applied cognitive science have the potential to offer, respectively, significant insights into learning and pedagogic practice.

We are also convinced, however, that the rapid popularisation of cognitive science inspired practice has led to the premature recommendation—and even mandating—of education practice underpinned by particular elements of cognitive science.

Of particular concern is the application of findings from particular subjects, age ranges, and contexts to other—often quite dissimilar—areas. Moreover, given the weaknesses in the applied evidence-base, cognitive science in the classroom is at present largely underpinned by evidence from controlled (laboratory) settings in conditions not typical of everyday classroom practice and with different populations such as university students. We suggest that the education community should not change its practices substantially without further applied evidence and more thorough and rigorous investigation into how practice might best be adapted.

Finally, our findings indicate that substantial investment is needed by the education profession to understand and model how practice might be adapted without eclipsing understandings of other important factors that influence learning, and ensure that members of the profession are skilled to understand and respond practically to these complexities.

C3. Limitations

In this section, we outline and briefly discuss the key limitations of this review's methodology and its implementation. We discuss limitations in the following areas:

- Focus, locating literature, and defining cognitive science informed strategies
- Configuration of strategy areas and groups
- Data extraction and quality appraisal
- Evidence synthesis
- Analysis of heterogeneity
- Discussions of theory, evidence, and practice
- Practice review sampling
- Conclusions and implications

Focus, locating literature, and defining cognitive science informed strategies

Have we defined and located the right bodies of literature? What has been emphasised? What has potentially been missed?

We defined the focus and scope of our review as follows:

> *This systematic review investigated approaches to teaching and learning informed by cognitive science that are commonly used in the classroom, with a particular focus on acquiring and retaining knowledge. This focus reflects the areas of cognitive science which have to date been the most influential for classroom practice and ostensibly have the most general application across the education sector (page 13).*

We are confident from our scoping work and the results from the practice review literature and data that we have successfully focused the review on cognitive science strategies that are *commonly used in the classroom.* One caveat here is that this potentially limits the boundaries of the review to that which has already been popularised. In many ways this is appropriate. We can be confident, due to the targeted searches for the focus cognitive science strategies, that we have identified the vast majority of relevant literature in all of our focus areas.

Where we are less confident, and where there are limitations for this study, are for strategies and areas that we have reviewed for which we did not conduct dedicated searches. The areas we have reviewed that were not included in our original focus strategies were Working with Schemas, Cognitive Theory of Multimedia Learning (to the extent to which this extends beyond Dual Coding, which was targeted), and Embodied Learning. We did not conduct dedicated searches for these wider areas and all material within them was located via searches for the target strategies and via general search terms for learning, memory, and cognitive science.[33] The Cognitive Theory of Multimedia Learning, and Embodied Learning, as terms in common use, are likely to have yielded more studies if dedicated searches had been conducted. This limitation also applies to some extent at the level of the specific strategies in our review areas. For example, we searched for terms relating to cognitive load and

[33] Our searches included the general cognitive science terms 'cognitive', 'brain', 'neuro', and 'learning science' and general memory terms such as 'working' and 'short-term' memory (related to dual coding and cognitive load, for example). See Appendix 3 for full details of the literature searches.

working memory but did not search specifically for 'worked examples'. Dedicated searches for specific practices may have revealed a wider evidence-base.

This challenge relates to a broader definitional one. We have set out to review strategies *informed by cognitive science,* rather than strategies that are in line with or can be linked to cognitive science. Within our priority criteria we looked for specific reference to cognitive science and the studies we have included can all be said to be informed by cognitive science. This is, however, a subset of practices in this area as there are many studies that employ (for example) concept mapping strategies without providing a rationale for doing so in terms of cognitive science. Our overall aim was to review the evidence for cognitive science informed practices; strictly, we were testing 'concept mapping' (for example) that is informed or motivated by cognitive science rather than concept mapping *per se*. If the focus of the review had been on the effectiveness of specific strategies *qua* strategies, more exhaustive searches would have been possible. However, this was beyond the scope of our review. Our strategies are instead representative of the value of applying principles from cognitive science to classroom practice rather than the focus in their own right.

Note that the problem here is twofold, relating to *literature location* and *definitions*. The definitional problem is that making this distinction was challenging and many of the studies were rated as medium or low relevance in the process, even when testing very similar strategies. The literature location problem is that our search terms were designed to locate the cognitive science practices rather than the strategy *per se*. While being less confident that our searches are exhaustive for the areas not identified as focus areas in the original searches, we maintain that our database is (a) unbiased and (b) comprehensive, even if it falls short of an exhaustive representation of the literature in the areas not specifically targeted. As per the distinction posed at the start of this section, if we solely focused on those areas that are already widespread in educational thinking, this may reinforce potential narrowness of scope. The lower confidence of full location of the literature in these areas is the trade-off for a more comprehensive coverage of the literature on 'cognitive science informed classroom practice'.

Configuration of strategy areas and groups

The organisation of studies into areas and strategies proved to be a highly complex and challenging aspect of the review. Categorisation was a two-stage process involving three researchers; several papers were moved between categories prior to them being finalised for analysis. At the point when we started analysis of the evidence, we had fixed the area-category-study groups in place. In many review areas, we noted how the studies linked to studies in other review areas. Many studies in our strategy groups were noted as not being entirely comparable during the analysis. There are several studies included in the main overview tables of results as being 'for information'. To maintain full transparency, we retained and reported these studies in the planned category, rather than move these while the analysis was underway. We hold that this was an unbiased process but the limitation here is that there are groups of studies that arguably should, or should not, have been grouped for analysis. Studies on interleaving may, for example, have been combined with studies of comparison and cognitive conflict to provide a larger weight of evidence in a slightly more general strategy area. In our analysis, we have discussed—to provide another example—worked examples and guidance and scaffolds alongside as ways of managing cognitive load and our overall conclusions in this area assume that there is sufficient similarity in the principles of these to make more general claims about the value of managing cognitive load.

Such decisions were at the level of areas, strategies, and individual studies in terms of what was included with what and where we drew the boundaries. We are of the view that this was an inevitable challenge with the nature of the literature and the general lack of shared, tight definitions and a common technical language for all the strategies. There were also issues at the level of the study, where strategies were not always described in detail and the resources available to the review did not allow for more extensive category configuration and analysis of strategy variants prior to the main analysis. In sum, the review has in many areas involved complex theoretical and definitional work to reach the areas and strategies in the final analysis. The limitation of this is that the judgement involved in this process lowers replicability and different researchers using different configurations could alter the emphasis and substantive features in the results.

Data extraction and quality appraisal

The main limitation in relation to data extraction is the level of detail that could be located in a database of this size within the review resources available. In each strategy area, we have provided a description of the evidence in terms of the populations represented, the subject or learning focus, the design and delivery of the intervention, and the outcome measures. We maintain that these are the most pertinent areas for focus, but note that (a) in many of these areas, more granular data would have been possible with more time and resource for the review, (b) we have inevitably made choices about which details to extract, and different choices might have added a different complexion to the characterisation and representativeness of the results, and (c) our consideration of these contextual factors was something we considered at the level of the strategy group. Further research might (for example, through meta-analytic and meta-regression techniques) bring these contextual factors or strategy variations into the analysis, analysing variation in the outcome estimates in relation to these factors. In contrast, our use of these variables was to assess applicability (for example, to note that studies were conducted in a limited number of subjects and therefore do not support generalisation beyond these) rather than to assess differential effectiveness (for example, that an effect is likely to be larger in subject A rather than B).

In the Protocol Implementation and Deviations section in Appendix 1, we report inter-rater reliability figures from the screening and eligibility assessment process. We judge this to be high, especially taking into account the complexity of the evidence and challenges of configuration discussed above. Below we discuss the synthesis and analysis of evidence for the high and medium priority studies. Before that, it is worth briefly commenting on why there was little value in low priority studies also being included within the analysis: as explained in our Methods sections (A3 and Appendix 1), these were largely a result of the two-step approach to eligibility assessment that applied broader criteria (for relevance and ecological validity) in a first step before a tighter second step. Low priority, in effect, makes a distinction between studies that are firmly excluded (failing one or more eligibility criteria categorically or by a wide margin) and those that were more marginally excluded. This adds a greater level of transparency to the review and was a response to some of the challenges of pre-specification of a complex and nascent evidence-base. Low priority studies were often very small (sometimes only a dozen pupils or less), of low ecological validity (nominally conducted in the classroom, but with little resemblance to everyday classroom conditions), and with only a weak adherence to the definitions of our focus cognitive science strategies.

Evidence synthesis

As we raise in our discussion of protocol implementation and deviations in Appendix 1, our approach to synthesis was refined after the initial screening and prior to analysis. We relied on an interpretation of the GRADE tool and a structured narrative summary as a key approach to maintain systematicity

and rigour, despite the high complexity and heterogeneity in the results. We hold that this approach successfully maintained systematicity and transparency in our synthesis. We recognise, however, that there is scope for judgement and disagreement in how the items assessed have led to (specific) overall judgements. Of utmost importance in the synthesis is that we have reported in detail every step of the analysis, and readers are in a position to follow and replicate our analysis.

Many of the limitations of our evidence synthesis stem back to the challenges of heterogeneity (discussed in relation to categorisation). We did not restrict the review to standardised programmes or near-identical practices as this would have resulted in the review being limited to only a very small number of studies. Equally, we were conscious of heterogeneity diluting, obscuring, and invalidating the findings. There was appreciable variation evident in relation to:

- the strategies in a given group;
- the populations and contexts in which they were tested;
- the choice of outcome measure;
- learning objective, topic, and subject area;
- comparisons and research questions tested;
- study size and quality;
- level of ecological validity (and therefore adherence to our eligibility criteria requiring tests in realistic classroom conditions);
- quality of the reporting; and
- relevance and adherence of studies to the (or our) definitions of the cognitive science areas and strategies.

This variation in the evidence presented practical and methodological challenges. The key practical challenge was the time and resource available to the review. In short, this presented a choice between the breadth and the depth and detail of the analysis. As discussed in relation to focus, we have prioritised coverage of cognitive science currently common in classroom practice. Given the large variation in classroom practice and our aim of aligning the review scope and content to prevailing practice, a broad and varied evidence-base was required. With greater resource and time, it would have been possible to (a) conduct a risk of bias assessment and (b) a more granular coding and data extraction along the lines of the variation described in the previous paragraph for all of our medium priority studies.

The lack of a risk of bias analysis for all studies in the analysis means that methodological limitations of studies included in the analysis may have been missed. We do not believe that this will have had a substantive effect on our analysis and findings. Our—already cautious—findings may have had confidence ratings (from GRADE) reduced further had more study limitations been identified. Our analysis has not, however, rested heavily on individual studies nor assumed that studies are low risk of bias until shown to be otherwise, and we have been cautious with the specificity of, and confidence in, our conclusions. The limitations in terms of cognitive science relevance, definition, and ecological validity for the medium eligibility group are, in our judgement, more significant considerations for the findings which could not have been apprised with a risk of bias analysis. The (adapted) GRADE assessment at the level of the group was more important for our findings.

Our main synthesis for each strategy group has followed an adaptation of the GRADE approach followed by a structured narrative summary of the evidence reporting (reporting the main finding, estimated impact, confidence in impact estimate, heterogeneity, and other points). The combination of these—an evidence group assessment and then summary—and transparent reporting of this represents a significant advance on narrative synthesis approaches typical of systematic reviews in

education research. Nonetheless, there is inevitably a substantial degree of subjectivity (that is, the use of expertise and judgement) involved in this approach. Judgement is required for all systematic reviews across all fields (Gough, Oliver and Thomas, 2017); cognitive science in the classroom and the complex, nascent, and often disparate nature of the applied evidence made this particularly so (see Discussions of Theory, Evidence, and Practice, below).

The GRADE approach is designed for assessment of quantitative evidence so the challenges of coding and analysing the impact of education interventions (a) required us to adapt the GRADE approach to appraise the evidence in more qualitative terms and (b) led us to prefer a structured narrative summary rather than a meta-analytic synthesis of the evidence for each group. As noted, the main limitation here relates to the practicalities of resources as well as the methodological questions this poses. Practically, it would have been possible to refine the GRADE assessment as well as the structured narrative summary with further quantitative work coding and analysing the data. Methodologically, the variation, gaps, vagueness, and subtlety of the evidence poses a danger that further quantification can obscure rather than refine the analysis. In other words, a detailed coding for quantitative analysis would have either ignored this heterogeneity and indirectness, presenting over-generalised and potentially misleading results—either that, or captured this complexity but thereby reduced cell counts and analytical groups to an unworkable number. In short, the overall limitation in the evidence and our approach led to more general and more subjective findings than we might have aspired to. The challenges outlined above aside, there *was* scope for refinement with greater resource and time. Nonetheless, we believe that the present analysis struck a good balance between several constraints and was well judged to provide the most informative, valid, and productive contribution to this developing field at this time.

Analysis of heterogeneity

We provide a summary of heterogeneity in our structured narrative synthesis based on descriptive data and GRADE analysis factors. It would have been possible with more time and resource to carry out a more granular coding and data extraction for all the medium eligibility studies. This would have fed into further analysis to examine how effectiveness varies by pupil groups and factors such as subject areas and strategy variation. This limitation, in our view, is more consequential than not conducting risk of bias analysis.

As we also discussed in the previous section, there are both practical and methodological problems with the quantification of an evidence-base of this nature. We discuss deviations from the original protocol in Appendix 1; a notable deviation and limitation relevant here is that in the original protocol we had planned to conduct more extensive analysis for subgroups of pupils (for example, primary and secondary), subject areas, and theoretically relevant strategy variations than was ultimately possible but, as noted, weaknesses in the evidence and the limitations of quantitative approaches to addressing variation formed the bases of the decision not to pursue heterogeneity analysis to the extent planned. As with the points on synthesis above, we believe that, while more refinement of the data was possible, our approach was justified given the evidence, objectives, and constraints. The limitation here is that our consideration of heterogeneity in the data has been more focused on identifying the gaps in the evidence rather than a break-down of the results by subgroups. The focus, when scrutinising variation, has been on identifying sufficient *homogeneity* for valid grouping of studies by strategy and identifying *gaps* in the subjects and populations represented—as opposed to scrutinising *within-group* variation to identify and enable subgroup analysis and analysis of variation in effects.

Discussions of theory, evidence, and practice

We have split each area for review into two main sections: first, the main systematic review of evidence and, second, an evidence-informed discussion and questions section. There are several possible limitations with this approach to dividing and reporting our data. The first criticism might be that the line we have drawn between (a) strategies for which there is a sufficient weight of evidence to be systematically reviewed and (b) those where the evidence must be reported as indicative wider evidence in the discussion section may be felt to be too strict or too loose. It is possible that some strategy areas might, with another review team or a different resource envelope, have been assigned to different sides of this divide.

An issue with the discussion and questions sections are that the evidential status of the various points made has been challenging to both determine and to communicate. In short, claims are made in the discussion sections that are (a) supported by *convincing indicative evidenc*e, (b) in line with, and supported by, *some evidence*, and (c) are plausible but that have *little evidential basis*: the claims, in other words, have varying levels of plausibility and explanatory power. Also, the extent to which evidence was *applied evidence* versus *basic scientific evidence* (but potentially of low ecological validity) was, again, hard to determine and communicate. It was not as simple as using all studies from our search database as ecologically-valid applied evidence and disregarding all else. Many of the studies in this section had been rated as *medium* ecological validity on our assessment tool. Many had a very strong basis in the basic science, others were 'second-hand' or, in our view, offered over-simplistic interpretations of the basic science. Throughout the discussions sections we have stressed the exploratory nature of the discussions and the uncertain and indicative nature of the propositions entertained and supported.

Practice review sampling

We have discussed the self-selected nature of the practice review interview and questionnaire sample in the practice review section (B9) and the associated Appendix (13). Sample bias is likely to be less problematic in the interviews due to the purposive sampling process. Our claims in relation the practice review data have deliberately avoided making claims about representativeness. Our discussion has been based on the assumption that these data sources reveal a large range of perspectives that are held but not that these are representative or held in any proportion in a given population or that these represent all possible perspectives.

Conclusions and implications

We have used systematic review tools to, as far as possible, ensure that conclusions follow in a transparent and valid manner from the evidence we have reviewed. There is inevitably some scope for judgement and disagreement in reaching overall statements of results. The process, being transparent and with pertinent study details reported, will allow readers to assess our conclusions and—where they hold alternative methodological principles or seek different standards of evidence—put them in a position to draw alternative conclusions. The final implications section is inevitably highly based on judgement and expertise. Again, we hold that these implications stem from the evidence we have presented and that readers can revisit the underlying evidence to reach different implications where they wish to apply alternative methodological principles or standards of evidence.

C4. About

Team

The review team brought together expertise in cognitive neuroscience, education policy and practice, education evaluation and measurement, systematic review, and interventions in the classroom. We drew together experts from the University of Birmingham, the Centre for the Use of Research and Evidence in Education (CUREE), and practitioner researchers from, or associated with, the Queen Anne's School and its neuroscience-focused research centre. Together the team offered a unique combination of expertise in undertaking systematic reviews of interventions in schools as well as knowledge of the underpinning science, the translation of new approaches in cognitive neuroscience to education, assessing the ecological validity of interventions, and the application of interventions in schools.

University of Birmingham

- ***Dr Thomas Perry*** led the project team. He has conducted numerous reviews including rapid, scoping, systematic, and policy reviews (Perry et al., 2018; Cordingley et al., 2018; Morris and Perry, 2017) and has specialist methodological expertise relating to research synthesis and review, quantitative methods and secondary data analysis, educational evaluation and improvement, social research methodology, and knowledge mobilisation, exchange, and use.
- ***Professor Deborah Youdell*** is an expert in the links between education policy and practice and inequalities. She is renowned for her development of interdisciplinary approaches in education drawing on new biosciences and neuroscience. She was Working Group Co-Chair of the UNESCO-funded International Science and Evidence based Education Assessment (ISEEA), bringing global experts together to map the state-of-the art for science-informed education.
- ***Professor Kimron Shapiro*** is an expert in cognitive neuroscience; in particular inattention, short- and long-term memory and their enhancement, as well as using modern approaches including electrophysiology and functional imaging to understand the brain mechanisms that underpin behaviour. Professor Shapiro has published over 100 papers, many in upper tier journals, and has an H-index of 42.
- ***Dr Rosanna Lea*** worked as a Research Fellow in the School of Education, University of Birmingham. Rose has substantial experience in conducting systematic reviews and her research interests include the psychology of education, applied cognitive science, and social and emotional learning.
- ***Dr Clara Rübner Jørgensen*** worked as a Research Fellow in the School of Education, University of Birmingham. Clara is a social anthropologist with experience of international and interdisciplinary research in education, extensive experience of conducting reviews, and expertise in involving stakeholders and communities in research and teaching.
- ***Niall Gamble*** worked as a Research Assistant in the School of Education, University of Birmingham. Niall is a postgraduate student at the University of Birmingham studying MSc Mental Health. He is interested in preventative and promotional strategies in mental health.
- ***Christina Pomareda*** worked as a Research Assistant in the School of Education, University of Birmingham. Christina is a postgraduate student at the University of Birmingham studying PhD Psychology. She has a keen interest in social cognition and individual differences.

Centre for the Use of Research and Evidence in Education (CUREE)

- ***Philippa Cordingley*** is an expert in systematic review, education evaluation, and education policy. She chaired the EPPI CPD Review group and has been Principal Investigator for three EPPI and numerous other full technical and less technical reviews.
- ***Paul Crisp*** is an expert in the technical aspects of knowledge management, design, analysis and interpretation, quantitative analysis, and reporting.

Queen Anne's School and BrainCanDo

- ***Julia Harrington***, ***Amy Fancourt***, *and colleagues at* ***Queen Anne's School*** are practitioner researchers with an expertise in using neuroscience research in practice in both state and independent schools. **BrainCanDo**[34] is an education neuroscience and cognitive psychology research centre based at Queen Anne's. Julie and colleagues supported the design, analysis, and reporting phases of the review, in particular in relation to review focus and usability. They were part of the advisory group and contributed to drawing out the implications of the results. Julia is Headmistress of Queen Anne's School and CEO of the BrainCanDo centre and Amy is Head of Psychology at QAS and Director of Research at BrainCanDo.

Advisory group

This research has benefited from and been developed in collaboration with a diverse advisory group which includes headteachers, cognitive neuroscientists, and experts in education research, policy, and practice. All members have a strong interest in the applications of cognitive science to educational contexts. The purpose of the advisory group was to contribute expertise relating to cognitive science, applied research, policy, and classroom practice. Furthermore, involving a wider team reduces bias when conducting systematic reviews (Utterly and Montgomery, 2017). The advisory group met three times during the project, at key points in the timeline, supplemented by ongoing opportunities for input via email communication. During these meetings, panel members provided their expertise and guidance on aspects of the project. The advisory group members were as follows:

- **Dr Robin Bevan**, Headteacher at Southend High School for Boys and National President of the National Education Union. He has a strong commitment to evidence-based practice, applying educational research into the classroom to enhance learning. He is a founding fellow of the Chartered College of Teaching.
- **Prof. Robert Coe** is Director of Research and Development at Evidence Based Education and Senior Associate at the Education Endowment Foundation. He was previously Professor of Education and Director of the Centre for Evaluation and Monitoring at Durham University, where he worked for 20 years doing research, evaluation, teaching and policy engagement. Rob is particularly interested in the uses of research by teachers and leaders and how it can be integrated into everyday pedagogy and school-level decision-making.
- **Dr Iroise Dumontheil** is a Reader in Cognitive Neuroscience, Birkbeck, Centre for Educational Neuroscience, on the Board of Directors for BrainCanDo, and advisory board for 'Learnus'. Her research focuses on social cognition and executive functions. She uses a variety of techniques, which include questionnaires, computerised tests, genetics, and structural and functional

[34] See https://braincando.com/

neuroimaging and is particularly interested in brain regions which support both social functions, such as mentalising or theory of mind, and executive functions such as relational reasoning, multitasking, prospective memory, and other cognitive processes such as mind-wandering or episodic memory retrieval.

- **Dr Amy Fancourt** is head of psychology at Queen Anne's School and director of research for BrainCanDo. She has expertise in, and experience of ,using neuroscience research in practice in both state and independent schools. She has published articles on psychology and education in journals, including *Nature Scientific Reports*, *Frontiers*, *Psychomusicology*, *Impact and Mind*, *Brain and Education*, and in media outlets, including *The Times*.
- **Dr Davinia Fernández-Espejo** is a Senior Lecturer at the University of Birmingham, School of Psychology. Her lab uses techniques such as MRI (structural and functional), tDCS, and behavioural approaches to test hypotheses about the role of different brain structures in cognition.
- **Julia Harrington** is headmistress at Queen Anne's School and founder and chief executive officer of BrainCanDo. She has featured in, and been interviewed by, publications and media outlets including *The Times*, the *Telegraph*, the *Guardian* and the BBC. Along with Amy and QAS colleagues, Julia has expertise in, and experience of, using neuro-science research in practice in both state and independent schools.
- **Niki Kaiser** is a chemistry teacher and network research lead at Notre Dame High School in Norwich. She is currently seconded part time to the Education Endowment Foundation as their science content specialist. In 2019, she won the Schools Education Award from the Royal Society of Chemistry for the incorporation and dissemination of research-informed teaching approaches.
- **Mark Stow** is Vice Principal and Director of Teaching and Learning at the University of Birmingham School. Mark joined the University of Birmingham School in January 2017. He is interested in evidence-informed approaches to teaching, and teacher development.
- **Prof. Hillevi Lenz Taguchi** is an expert in Child and Youth Studies, and Early Childhood Education currently working on transdisciplinary studies in Learning-Brain-Practice in preschool. This particular research compares and tests how socio-emotional learning practices and computer programme-oriented practices in Swedish preschools affect children's attention, social understanding, language, and communication skills. This interdisciplinary project incorporates methodologies from pedagogy, language skill estimations from linguistics, as well as cortical measurements from cognitive neuroscience.
- **Sonia Thompson**, Headteacher and Research School Director of St. Matthew's Church of England teaching and research school, Birmingham. Sonia is an SLE for English and School Improvement, Accredited Talk for Writing Training Centre Lead and a former English Consultant for Birmingham LA. She has had articles published in the *Chartered College of Teaching Impact journal* and *Schools Week*.
- **Prof. Sam Twiselton** is the Director of Sheffield Institute of Education at Sheffield Hallam University and Vice President (External) of the Chartered College of Teaching. She uses her research and practice in the development of teacher expertise and curriculum design to develop approaches to teacher development. Sam has been involved in influencing government policy on teacher education and as the Chair of the DfE ITT Framework Group, a member of the DfE Teacher Recruitment and Retention Advisory Group, the specialist NPQs Group, the Carter Review of ITT and Expert Behaviour Management Panel, and the OFSTED Curriculum Review Panel. She is a recent recipient of an OBE for services to higher education.

Conflicts of interest

All members of the research team, as listed above, are committed to educational practice being informed, where appropriate, by cognitive science, and the development, configuration and application of cognitive science to that end. We are also committed to a disinterested, systematic, and transparent review processes to ensure that our assessment of the evidence-base is warranted on methodological grounds and open to scrutiny. We have no conflicts of interest in our ability or intention to carry out these commitments. Nonetheless, several members of the team were/are working on related projects, summarised below:

- Youdell was Co-Chair of UNESCO International Science and Evidence in Education Assessment.
- Shapiro was PI on a research programme to test the application of transcranial stimulation (tCS) to improve working memory. At the time of the research, he was seeking a patent and in discussion with various companies about commercialisation.
- Harrington is the CEO and Fancourt is the Director of Research for BrainCanDo, a charitable company whose aims include the development of a strong neuroscientific evidence-base to inform and underpin education. BrainCanDo aims to empower teaching professionals to use the latest findings from neuroscience research to transform and enrich their classrooms and to empower students to understand how learning happens. One of the programmes that they trialled in 2020 was a 'Neuroscience for Teachers' course developed by Prof. Patricia Riddell. BrainCanDo also published a book in July 2020: *The Braincando Handbook of Teaching and Learning: Practical Strategies to Bring Psychology and Neuroscience Into the Classroom*.

The research was organised in terms of a core team (TP, RL, CJ) who conducted the searches, screening, and data extraction, including the study eligibility criteria and quality appraisal coding. The wider team were involved in review design, analysis, and narrative synthesis; NG and CP supported during finalising the study database and report.

Registration

This systematic review was registered on the Open Science Framework registry following publication of the protocol. The details of this are provided below:

Description: Systematic Review of applications of Cog Sci in the R-18 classroom commissioned by the Education Endowment Foundation
Category: Project **Registration type:** Open-Ended Registration
Date registered: December 22, 2020
URL: https://osf.io/y839t/

The protocol is publicly available on the EEF website: https://educationendowmentfoundation.org.uk/evidence-summaries/evidence-reviews/cognitive-science-approaches-in-the-classroom/

C5. Report references

Agarwal, P. K. and Bain, P. M. (2019) *Powerful Teaching: Unleash the Science of Learning*, San Francisco: Jossey-Bass.

Ainscow, M., Chapman, C. and Hadfield, M. (2019) *Changing Education Systems: A Research-Based Approach*, Routledge.

Alferink, L. A. and Farmer-Dougan, V. (2010) 'Brain-(Not) Based Education: Dangers of Misunderstanding and Misapplication of Neuroscience Research', *Exceptionality*, 18 (1), pp. 42–52.

Amin, H. U., Malik, A. S. and Kamel, N. (2014) 'Memory Retention and Recall Process', in *EEG/ERP Analysis: Methods and Applications*, CRC Press (pp. 219–237).

Aronsson, L. and Lenz Taguchi, H. (2017) 'Mapping a Collaborative Cartography of the Encounters Between the Neurosciences and Early Childhood Education Practices', *Discourse Studies in the Cultural Politics of Education,* 39 (2). pp. 1–16. DOI: 10.1080/01596306.2017.1396732

Atkinson, R. C. and Shiffrin, R. M. (1968) 'Human Memory: A proposed System and Its Control Processes', in *Psychology of Learning and Motivation,* Academic Press (vol. 2, pp. 89–195).

Arwood, E. L. and Meridith, C. (2017) *Neuro-Education: A Translation from Theory to Practice*, Tigard, Oregon: Arwood Neuro-Viconics.

Baddeley, A. D. and Hitch, G. (1974) 'Working Memory', in *Psychology of Learning and Motivation*, Academic Press (vol. 8, pp. 47–89).

Bevilacqua, D., Davidesco, I., Wan, L., Chaloner, K., Rowland, J., Ding, M., . . . Dikker, S. (2018) 'Brain-to-Brain Synchrony and Learning Outcomes Vary by Student–Teacher Dynamics: Evidence from a Real-world Classroom Electroencephalography Study', *Journal of Cognitive Neuroscience*, 31 (3), pp. 401–411. DOI: 10.1162/jocn_a_01274

Bjork, R. A. and Bjork, E. L. (1992) 'A New Theory of Disuse and an Old Theory of Stimulus Fluctuation', in *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes*, 2, pp. 35–67.

Brown, P. C., Roediger III, H. L. and McDaniel, M. A. (2014) *Make It Stick*, Harvard University Press.

Bryant-Khachy, F. (2018a, June) 'A Randomised Controlled Trial of Spaced Learning in a Primary School Context. Study A: KS1 Parallel Replication (Geography)' [Year Group Replications Reported: Y1, Y2], poster session presented at Academy of Principals, 9th Global Educational Leadership Conference, Singapore.

Bryant-Khachy, F. (2018b, June). 'A Randomised Controlled Trial of Spaced Learning in a Primary School Context. Study B: KS2 Parallel Replication (History)' [Year Group Replications Reported: Y3, Y4, Y5, Y6], poster session presented at Academy of Principals, 9th Global Educational Leadership Conference, Singapore.

Borenstein, M., Hedges, L. V., Higgins, J. P. T. and Rothstein, H. R. (2009) *Introduction to Meta-Analysis*, Wiley.

Buzsák, G. (1998) 'Memory Consolidation During Sleep: A Neurophysiological Perspective', *Journal of Sleep Research*, 7 (S1), pp. 17–23.

Callan, D. E. and Schweighofer, N. (2010) 'Neural Correlates of the Spacing Effect in Explicit Verbal Semantic Encoding Support the Deficient-Processing Theory', *Human Brain Mapping*, 31 (4), pp. 645–659.

Carpenter, S. K. and Agarwal, P. K. (2020) 'How to Use Spaced Retrieval Practice to Boost Learning', Iowa State University.

Carpenter, S. K. (2009) 'Cue Strength as a Moderator of the Testing Effect: The Benefits of Elaborative Retrieval', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35 (6), p. 1563.

Caviglioli, O. (2019) *Dual Coding with Teachers*, John Catt Educational Limited.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T. and Pashler, H. (2008) 'Spacing Effects in Learning: A Temporal Ridgeline of Optimal Retention', *Psychological Science*, 19 (11), pp. 1095–1102.

Clouter, A., Shapiro, K. L. and Hanslmayr, S. (2017) 'Theta Phase Synchronization Is the Glue that Binds Human Associative Memory', *Current Biology*, 27, pp. 1–6.

Collins, S. (2019) *Neuroscience for Learning and Development: How to Apply Neuroscience and Psychology for Improved Learning and Training*, Kogan Page.

Constantinidis, C. and Klingberg, T. (2016) 'The Neuroscience of Working Memory Capacity and Training', *Nature Reviews Neuroscience*, 17 (7), p. 438.

Cordingley, P., Greany, T., Crisp, B., Seleznyov, S., Bradbury, M. and Perry, T. (2018) 'Developing Great Subject Teaching: Rapid Evidence Review of Subject-Specific Continuing Professional Development in the UK', Wellcome Trust: http://www.curee.co.uk/node/5032

Cowan N. (2014) 'Working Memory Underpins Cognitive Development, Learning, and Education', *Educational Psychology Review*, 26 (2), pp. 197–223. https://doi.org/10.1007/s10648–013–9246-y

Cowan, N. (2016) '*Working Memory Capacity' (Classic Edition)*, Psychology Press.

Davidesco, I., Laurent, E., Valk, H., West, T., Dikker, S., Milne, C. and Poeppel, D. (2019) 'Brain-to-Brain Synchrony Predicts Long-Term Memory Retention More Accurately Than Individual Brain Measures', *bioRxiv*, 644047. DOI: 10.1101/644047

Deans for Impact (2015) 'The Science of Learning': Deans for Impact: http://www.deansforimpact.org/wp-content/uploads/2016/12/The_Science_of_Learning.pdf

Dehaene, S. (2020) *How We Learn: The New Science of Education and the Brain*, Penguin UK.

De Jong, T. (2010) 'Cognitive Load Theory, Educational Research, and Instructional Design: Some Food for Thought', *Instructional Science*, 38, pp. 105–134. https://doi.org/10.1007/s11251–009–9110–0

Didau, D. and Rose, N. (2016) *What Every Teacher Needs to Know about... Psychology*, John Catt Educational.

Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., . . . Poeppel, D. (2017) 'Brain-to-Brain Synchrony Tracks Real-World Dynamic Group Interactions in the Classroom', *Current Biology*, 27 (9), pp. 1375–1380. DOI: 10.1016/j.cub.2017.04.002

Dunlosky, J. and Rawson, K. (2015) 'Practice Tests, Spaced Practice, and Successive Relearning: Tips for Classroom Use and for Guiding Students' Learning', *Scholarship of Teaching and Learning in Psychology*, 1, pp. 72–78.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J. and Willingham, D. T. (2013) 'Improving Students' Learning with Effective Learning Techniques: Promising Directions from Cognitive and Educational Psychology', *Psychological Science in the Public Interest*, 14 (1), pp. 4–58.

Enser, Z. and Enser, M. (2020) *Fiorella and Mayer's Generative Learning in Action*, John Catt Educational.

Fiorella, L. and Mayer, R. E. (2016) 'Eight Ways to Promote Generative Learning, *Educational Psychology Review*, 28 (4), pp. 717–741.

Fiorella, L. and Mayer, R. E. (2015) *Learning as a Generative Activity*, Cambridge University Press.

Fischer, K. W., Goswami, U. and Geake, J. (2010) 'The Future of Educational Neuroscience. *Mind, Brain, and Education*, 4 (2), pp. 68–80.

Gluckman, M., Vlach, H. A. and Sandhofer, C. M. (2014) 'Spacing Simultaneously Promotes Multiple Forms of Learning in Children's Science Curriculum', *Applied Cognitive Psychology*, 28 (2), pp. 266–273.

Goswami, U. (2006) 'Neuroscience and Education: From Research to Practice?', *National Review of Neuroscience*, 7(5), pp. 406–411

Goswami, U. (2015) 'The Neural Basis of Dyslexia May Originate in Primary Auditory Cortex', *Brain*, 137 (12), pp. 3100–3102. DOI: 10.1093/brain/awu296.

Gough, D., Oliver, S. and Thomas, J. (eds) (2017) *An Introduction to Systematic Reviews*, Sage.

Harrington, J., Beale, J., Fancourt, A. and Lutz, C. (eds) (2020) *The 'BrainCanDo' Handbook of Teaching and Learning: Practical Strategies to Bring Psychology and Neuroscience Into the Classroom*, Routledge.

Howard-Jones, P. (2014) 'Neuroscience and Education: Myths and Messages', *Nature Reviews Neuroscience*, 15, pp. 817–824. DOI: 10.1038/nrn3817

Howard-Jones, P., Ioannou, K., Bailey, R., Prior, J., Hui Yau, S. and Jay, T. (2018) 'Applying the Science of Learning in the Classroom', *Impact*: https://impact.chartered.college/article/howard-jones-applying-science-learning-classroom/

Ionescu, T. and Vasc, D. (2014) 'Embodied Cognition: Challenges for Psychology and Education', *Procedia-Social and Behavioral Sciences*, 128, pp. 275–280.

Jensen, E. and McConchie, L. (2020) *Brain-Based Learning: Teaching the Way Students Really Learn*, Corwin.

Jones, K. (2019) *Retrieval Practice: Research and Resources for Every Classroom*, John Catt Educational.

Jonsson, B., Wiklund-Hörnqvist, C., Stenlund, T., Andersson, M. and Nyberg, L. (2020) 'A Learning Method for All: The Testing Effect Is Independent of Cognitive Ability', *Journal of Educational Psychology*. Advance online publication: http://dx.doi.org/10.1037/edu0000627

Kang, S. H., Lindsey, R. V., Mozer, M. C. and Pashler, H. (2014) 'Retrieval Practice Over the Long Term: Should Spacing Be Expanding or Equal-Interval?', *Psychonomic Bulletin & Review*, 21 (6), pp. 1544–1550.

Kang, S. H. K. and Pashler, H. (2012) 'Learning Painting Styles: Spacing is Advantageous When It Promotes Discriminative Contrast', *Applied Cognitive Psychology*, 26, pp. 97–103.

Kirschner, P. A. and Hendrick, C. (2020) *How Learning Happens: Seminal Works in Educational Psychology and What They Mean in Practice*, Routledge.

Kroeger, L. A., Douglas Brown, R. and O'Brien, B. A. (2012) 'Connecting Neuroscience, Cognitive, and Educational Theories and Research to Practice: A Review of Mathematics Intervention Programs, *Early Education and Development*, 23 (1), pp. 37–58.

Küpper-Tetzel, C. E., Erdfelder, E. and Dickhäuser, O. (2014) 'The Lag Effect in Secondary School Classrooms: Enhancing Students' Memory for Vocabulary', *Instructional Science*, 42 (3), pp. 373–388.

Kyle, F., Kujala, J. V., Richardson, U., Lyytinen, H. and Goswami, U. (2013) 'Assessing the Effectiveness of Two Theoretically Motivated Computer-Assisted Reading Interventions in the United Kingdom: GG Rime and GG Phoneme', *Reading Research Quarterly*, 48 (1), pp. 61–76.

Lakoff, G. (2015) 'How Brains Think: The Embodiment Hypothesis', in keynote address recorded March 14, 2015 at the inaugural International Convention of Psychological Science.

Lavie, N. and De Fockert, J. (2005) 'The Role of Working Memory in Attentional Capture', *Psychonomic Bulletin & Review*, 12 (4), pp. 669–674.

Ling Lo, M. (2012) *Variation Theory and the Improvement of Teaching and Learning*, Göteborg: Acta Universitatis Gothoburgensis.

Lovell, O. (2020) *Sweller's Cognitive Load Theory in Action*, John Catt Educational.

MacCann, C., Jiang, Y., Brown, L., Double, K., Bucich, M. and Minbashian, A. (2020) 'Emotional Intelligence Predicts Academic Performance: A Meta-Analysis', *Psychological Bulletin*, 146 (2), pp. 150–186. https://doi.org/10.1037/bul0000219

Mason, A., Farrell, S., Howard-Jones, P. and Ludwig, C. J. H. (2017) 'The Role of Reward and Reward Uncertainty in Episodic Memory', *Journal of Memory and Language*, 96, pp. 62–77. https://doi.org/10.1016/j.jml.2017.05.003

Mayer, R. E. (2021) *Multimedia Learning (3rd Edition),* Cambridge University Press.

Mayer, R. E. (2005) 'Cognitive Theory of Multimedia Learning', in *The Cambridge Handbook of Multimedia Learning*, 41, pp. 31–48.

Mayer, R. E. (2002) 'Multimedia Learning', in *Psychology of Learning and Motivation*, Academic Press (41, pp. 85–139).

Mayer, R. E. and Anderson, R. B. (1991) 'Animations Need Narrations: An Experimental Test of a Dual-Coding Hypothesis', *Journal of Educational Psychology*, 83 (4), p. 484.

Medina, J. (2008) *Brain Rules: 12 Principles for Surviving and Thriving at Work, Home and School*, Seattle, Washington: Pear Press.

Mitchell, C., Nash, S. and Hall, G. (2008) 'The Intermixed-Blocked Effect in Human Perceptual Learning is Not the Consequence of Trial Spacing', Jo*urnal of Experimental Psychology: Learning, Memory, and Cognition,* 34, pp. 237–242.

Morris, R. and Perry, T. (2017) 'Reframing the English Grammar Schools Debate', *Educational Review*: http://dx.doi.org/10.1080/00131911.2016.1184132.

Muijs, D. and Bokhove, C. (2020) 'Metacognition and Self-Regulation: Evidence Review', London: Education Endowment Foundation: https://educationendowmentfoundation.org.uk/evidence-summaries/evidence-reviews/metacognition-and-self-regulation-review/.

Murre, J. M. and Dros, J. (2015) 'Replication and Analysis of Ebbinghaus' Forgetting Curve', *PloS One*, 10 (7), e0120644.

Ofsted (2019) 'Education Inspection Framework: Overview of Research': https://www.gov.uk/government/publications/education-inspection-framework-overview-of-research

Pan, S. and Agarwal, P. (2018) 'Retrieval Practice and Transfer of Learning: Fostering Students' Application of Knowledge': http://retrievalpractice.org

Paas, F. and Sweller, J. (2012) 'An Evolutionary Upgrade of Cognitive Load Theory: Using the Human Motor System and Collaboration to Support the Learning of Complex Cognitive Tasks', *Educational Psychology Review*, 24 (1), pp. 27–45.

Paivio, A. (1991) 'Dual Coding Theory: Retrospect and Current Status', *Canadian Journal of Psychology/Revue Canadienne De Psychologie*, 45 (3), p. 255.

Perry, T., Cordingley, P., Johns, P. and Bradbury, M. (2018) 'International Review of Teacher Evaluation Systems: Executive Summary, Main Report, Technical Report and System Case Studies', prepared for the Inter-American Development Bank (IDB).

Plass, J. and Kalyuga, S. (2019) 'Four Ways of Considering Emotion in Cognitive Load Theory', *Educational Psychology Review*, 31, pp. 339–359.

Potts, R. and Shanks, D. R. (2014) 'The Benefit of Generating Errors During Learning', *Journal of Experimental Psychology: General*, 143 (2), p. 644; and, more generally, Brown, P. C., Roediger III, H. L. and McDaniel, M. A. (2014) *Make it Stick*, Harvard University Press.

Purdy, N. (2008) 'Neuroscience and Education: How Best to Filter Out the Neurononsense from Our Classrooms?', *Irish Educational Studies*, 27 (3), pp. 197–208.

Putnam, A. L. and Roediger III, H. L. (2018) 'Education and Memory: Seven Ways the Science of Memory Can Improve Classroom Learning', in *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, 1, pp. 1–45.

Roediger III, H. L., Putnam, A. L. and Smith, M. A. (2011) 'Ten Benefits of Testing and Their Applications to Educational Practice', *Psychology of Learning and Motivation*, 55, pp. 1–36.

Rohrer, D., Dedrick, R. F. and Burgess, K. (2014) 'The Benefit of Interleaved Mathematics Practice is Not Limited to Superficially Similar Kinds of Problems', *Psychonomic Bulletin & Review*, 21 (5), pp. 1323–1330.

Rose, S. and Rose, H. (2013) *Genes, Cells and Brains: The Promethean Promise of the New Biology*, London: Verso.

Sawyer, R. K. (2006) 'The New Science of Learning' in *The Cambridge Handbook of the Learning Sciences*, 1, p. 18.

Schlesinger, M. A., Hassinger-Das, B., Zosh, J. M., Sawyer, J., Evans, N. and Hirsh, P. (2020) 'Cognitive Behavioral Science behind the Value of Play: Leveraging Everyday Experiences to Promote Play, Learning, and Positive Interactions', *Journal of Infant, Child and Adolescent Psychotherapy*, 19 (2): https://doi.org/10.1080/15289168.2020.1755084

Shapiro, L. (2019) *Embodied Cognition*, Routledge.

Sherrington, T. and Caviglioli, O. (2020) *Teaching Walkthrus: Five-Step Guides to Instructional Coaching*, John Catt Educational.

Smith (n.d.) 'When is a Chunk Not a Chunk?' (blog): https://theemotionallearner.com/2019/10/25/when-is-a-chunk-not-a-chunk/

Smolen, P., Zhang, Y. and Byrne, J. (2016) 'The Right Time to Learn: Mechanisms and Optimization of Spaced Learning', *Nature Review Neuroscience*, 17 (2), pp. 77–88.

Soderstrom, N. C. and Bjork, R. A. (2015) 'Learning Versus Performance: An Integrative Review', *Perspectives on Psychological Science*, 10 (2), pp. 176–199.

Sweller, J. (1988) 'Cognitive Load During Problem Solving: Effects on Learning', *Cognitive Science*, 12 (2), pp. 257–285.

Sweller, J. (1994) 'Cognitive Load Theory, Learning Difficulty, And Instructional Design', *Learning and Instruction*, 4, pp. 295–312.

Sweller, J., van Merriënboer, J. J. and Paas, F. (2019) 'Cognitive Architecture and Instructional Design: 20 Years Later', *Educational Psychology Review*, 31 (2), pp. 261–292.

Sweller, J. (2016) 'Cognitive Load Theory, Evolutionary Educational Psychology, and Instructional Design', in Geary, D. C. and Berch, D. B. (eds), *Evolutionary Perspectives on Child Development and Education*, Springer (pp. 291–306).

Tibke, J. (2019) *Why the Brain Matters: A Teacher Explores Neuroscience*, SAGE.

Utterly, L. and Montgomery, P. (2017) 'The Influence of the Team in Conducting a Systematic Review', *Systematic Review*, 6 (1): https://doi.org/10.1186/s13643–017–0548-x

Vekiri, I. (2002) 'What is the Value of Graphical Displays in Learning?', *Educational Psychology Review*, 14 (3), pp. 261–312.

Von Bastian, C. C. and Oberauer, K. (2014) 'Effects and Mechanisms of Working Memory Training: A Review', *Psychological Research*, 78 (6), pp. 803–820.

Yilmaz, K. (2011) 'The Cognitive Perspective on Learning: Its Theoretical Underpinnings and Implications for Classroom Practices', *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 84 (5), pp. 204–212

Yang, C., Luo, L., Vadillo, M. A., Yu, R. and Shanks, D. R. (2021) 'Testing (Quizzing) Boosts Classroom Learning: A Systematic and Meta-Analytic Review', *Psychological Bulletin*, 147 (4), pp. 399–435. https://doi.org/10.1037/bul0000309

Youdell, D. and Lindley, M. R. (2018) *Biosocial Education: The Social and Biological Entanglements of Learning*, London: Routledge.

Weinstein, Y., Madan, C. R. and Sumeracki, M. A. (2018) 'Teaching The Science of Learning', *Cognitive Research: Principles and Implications*, 3 (1), p. 2.

Weinstein, Y., Sumeracki, M. and Caviglioli, O. (2018) *Understanding How We Learn: A Visual Guide*', Routledge.

Wigelsworth, M., Verity, L., Mason, C., Humphrey, N., Qualter. P. and Troncoso, P. (2020) 'Programmes to Practices: Identifying Effective, Evidence-Based Social and Emotional Learning Strategies for Teachers and Schools: Evidence Review', London: Education Endowment Foundation: https://educationendowmentfoundation.org.uk/evidence-summaries/evidence-reviews/social-and-emotional-learning/

Wiklund-Hörnqvist, C., Stillesjö, S., Andersson, M., Jonsson, B. and Nyberg, L. (2021) 'Retrieval Practice Facilitates Learning by Strengthening Processing in Both the Anterior and Posterior Hippocampus', *Brain and Behavior*, 11:e01909: https://doi.org/10.1002/brb3.1909

Wiliam, D. (2016) *Leadership for Teacher Learning*, West Palm Beach, FL: Learning Sciences International.

Willis, J. (2009) ‘What Brain Research Suggests for Teaching Reading Strategies’, *The Educational Forum*, 73 (4), pp. 333–346. DOI: 10.1080/00131720903166861

Wilson, M. (2002) ‘Six Views of Embodied Cognition’, *Psychonomic Bulletin & Review*, 9 (4), pp. 625–636.

Worth, J., Nelson, J., Harland, J., Bernardinelli, D. and Styles, B. (2018) ‘GraphoGame Rime: Evaluation Report and Executive Summary’, London: Education Endowment Foundation.

Appendices

Appendix 1: methodology: systematic review

The following information is from the review protocol. This information is copied directly and unchanged for transparency purposes. Following the description of the methods, there is a short 'Protocol Implementation and Deviations' section reporting information about conduct of the review, including reporting of inter-rater reliability data, methods decisions made during implementation and minor differences between the protocol plans and the details of how they were implemented in the review.

Inclusion and exclusion criteria for the review

Our approach to searching and screening is iterative, with two overall groups of eligibility criteria to be applied. The initial eligibility criteria for studies to include in the core systematic review, as per the review protocol, are the following:

1. **Population:** Children and young people between 3 and 18 years of age in classroom settings (i.e., excluding University based studies with first year undergraduate students). This can include mainstream or special education settings, in any country. While schools are generally attended by children aged five upwards, a growing body of evidence suggests that cognitive science is useful for helping understand how play contributes to learning processes in very young children (e.g., see Schlesinger *et al*., 2020). Although the 'classroom' may not necessarily be in a school, learning can still take place in other Early Years settings (e.g., playgroup). Furthermore, recent published EEF reviews on other aspects of learning, such as metacognition and self-regulated learning (Muijs & Bokhove, 2020), and social and emotional learning (Wigelsworth *et al.*, 2020), have also included studies with pre-school age children (i.e., 3-4 years) and adolescents up to age 18 (i.e., students in Further Education). Including a broad age range will ensure our review aligns with those other EEF reviews.
2. **Interventions/Practices of interest**:
    i. An evaluation of a classroom trials and/or interventions
    ii. Uses approaches derived from cognitive science relating to the acquisition and retention of knowledge. Studies need to demonstrate that the approach is derived from or inspired by cognitive science theory/principles. Scoping for each of the key areas outlined in the concept map (Appendix 2) will help outline the key terminology, theories and underpinning science for that specific area, which will then be used as a basis to judge whether a strategy is sufficiently based on cognitive science. For example, a classroom study that includes reference to 'germane load' or cites the work of Sweller, has clearly been informed by cognitive load theory. If a study is excluded as not overtly being derived from cognitive science, there may still be rationale for it to be included in the practice review to explore practices related to those that are.
3. **Study design and outcomes:**
    i. Initially we will include all studies reporting empirical evidence of any type or quality about pupil impact.
    ii. At the initial screening stage, we will include all studies that have *any form* of counterfactual (i.e., within- or between-group comparison of outcomes between

conditions), even where serious threats to internal validity, such as selection bias, contamination or group imbalance are evident or probable.

Following an appraisal of the coverage and quality of evidence (see below) across the various cognitive science areas, we expect to tighten this criterion to include only experimental and quasi-experimental studies, which we will code using the EEF data extraction framework (v1.0, October 2019) (please see below for details of our code sets/extraction frameworks). In the former, we include randomised experiments (randomised at any level). In the latter we will include (in addition to the EEF framework) well-controlled observational/correlational designs such as instrumental variables regression approaches, prospective and retrospective propensity score matching, difference-in-difference methods and regression discontinuity designs (or similar, such as interrupted time series from natural experiments) (see Shadish, Cook & Campbell, 2002). We also expect to tighten this criterion to include only studies reporting test-based outcomes (measured in either grades and/or cognitive outcomes). We have included this initial 'low bar' for eligibility screening to cover the possibility that the amount and quality of evidence in a given area of cognitive science (as per our concept map) is limited, in which case we would retain studies providing only low internal validity evidence (which we will report as such). The approach to quality appraisal is set out in detail below.

iii. Systematic reviews and meta-analyses will also be included in initial searches from which empirical studies will be mined and assessed against these criteria. Results and conclusions from reviews will not be included in the core systematic review but will inform the review sub-strands and thereby be a point of comparison during review analysis and synthesis. We will flag (but exclude) all reviews as such in initial screening using an eligibility code.

iv. Studies that do not meet the eligibility criteria for the core review may still be included in either the practice review or the underpinning science review where they address the questions and objectives of these sub-strands, as specified above. These will be flagged (but excluded)

4. **Language:** written in English and peer reviewed (for journal articles).
5. **Date:** All
6. **Bodies of Literature, included in the review:**
    i. All peer reviewed journal articles. To limit search results, we will filter out non-peer reviewed journal articles at the search stage, where possible. Where databases do not include a filter for peer reviewed articles, we will refer to the journals themselves for information.
    ii. Reports based on research commissioned by policy makers, charitable or other non-commercial organisations.
    iii. Due to the time and cost constraints for the review, we will exclude conference proceedings, working papers and master's and doctoral dissertations/theses *that were published before January 2017* on the grounds that high-quality studies are likely to have been subsequently published.

It should be emphasised that the above describes the initial screen. We detail the iterative process through which further criteria are applied for inclusion and extraction in more detail below.

Search strategy for identification of studies

Search Databases

For the review we will search the following databases, which include generic and specialist material:

Table X1.1 – Search Databases

<table>
<tr><th>Databases Included from the Web of Science Core Collection</th><th>Databases Included from the ProQuest Collection</th></tr>
<tr><td>• Science Citation Index Expanded (SCI-EXPANDED)
• Social Sciences Citation Index (SSCI)
• Arts & Humanities Citation Index (A&HCI)
• Book Citation Index– Science (BKCI-S)
• Book Citation Index– Social Sciences & Humanities (BKCI-SSH)
• Emerging Sources Citation Index (ESCI)</td><td>• Ebook Central
• Education Database
• Psychology Database
• Social Science Database
• Sociology Database
• ERIC
• International Bibliography of the Social Sciences (IBSS)
• Applied Social Sciences Index & Abstracts (ASSIA)
• Sociological Abstracts</td></tr>
<tr><th>Science Direct</th><th rowspan="3">JSTOR
For titles in the following areas:
• Education
• General Science
• Psychology
• Science and technology studies</th></tr>
<tr><td>All publications</td></tr>
<tr><th>EBSCO</th></tr>
<tr><td>Education databases</td><td></td></tr>
</table>

We will also search in the following databases:

- **Campbell Collaboration Library of Systematic reviews:**
  - https://onlinelibrary.wiley.com/journal/18911803
- **Cochrane Library of Systematic Reviews:**
  - https://www.cochranelibrary.com/search?cookiesEnabled
- **EPPI Centre library of reviews:**
  - https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=62
- **Open Science Framework:**
  - https://osf.io/registries?view_only=
- **Google Scholar**
  - To capture commissioned research reports, we will also search Google Scholar and review the first 1000 results.

Additional searches:

- Two high impact and relevant journals selected for handsearching: ***Trends in Neuroscience*** and ***Education and Educational Psychology Review***. The reference lists for included studies and relevant systematic reviews will also be manually searched.

We **do not** plan to search in the following databases as we feel they are of lower relevant to our focus and/or will have too much overlap with the above to be effective use of the review resources:

- PubMed
- FirstSearch

Following searches, we will also liaise with EEF colleagues to identify relevant studies already in the EEF Education Database. An application will be submitted to the EEF to access these studies. We will add these to the main results and retaining records for which data are already extracted when removing duplicates.

Where databases allow screening by category (e.g., Web of Science, JSTOR), this will be used to remove records which are not relevant at the initial search stage.

We will be using the EPPI-Reviewer software for the review (see below for further details).

Search Terms

Our search strings have been developed through preliminary database searches to assess search term sensitivity and precision.[35] We also have considered feedback from advisory group members (see Appendix 1) about where to prioritise and how to define cognitive science concepts. The search terms will be based on the interventions outlined in our conceptual map (see Appendix 2 and 3).

For each concept, the string will contain terms related to a) methodology, b) education (outcomes and classroom specific), and c) terms and synonyms related to the specific cognitive science area (including a general cognitive science search). These search terms will be entered into each search database with the minimum of adaptation needed to use the search syntax and functionality and ensure comparability across databases.

Table X1.2 – General Search Terms (all searches)

Search Term Group	***Search String (Fragment)***	***Search Location[1]***
Group 1 – Methodology	intervention OR trial OR evaluat* OR experiment* OR quasi-experiment* OR pilot OR test*	Title, abstract or key words
Group 2 – Education Outcomes	AND learning OR attainment OR achievement OR "test scores" OR outcomes OR exam* OR impact OR effect OR performance	
Group 3 – Classroom setting	AND classroom OR teach* OR school OR “further education” OR nursery OR "early years" OR kindergarten OR pre-primary OR lesson	
Group 4 – Focus Concept	AND, one of the general or concept-specific search term fragments in Table 4, below.	

[1] Subject to search database functionality

The general search terms above will be combined with one of the search strings related to cognitive science in general and specific cognitive science concepts, below.

[35] https://handbook-5-1.cochrane.org/chapter_6/6_4_4_sensitivity_versus_precision.htm

Table X1.3 – Cognitive Science Concept-specific Search Terms – Core Concepts

Cognitive Science Concept	***Search String (Fragment – to be combined with the general search terms, above)***	***Search Location[1]***
Cognitive Science General	cog* OR brain* OR neuro* OR "learning science"	Title, abstract or key words
Spaced practice	spac* OR distributed	
Interleaving	interleav* OR interweav*	
Retrieval practice	retriev* OR "testing effect"	
Dual coding	dual	
Strategies to manage cognitive load	"working memory" OR "short-term memory" OR (load AND (Cognitive OR intrinsic OR extraneous OR germane))	

[1] Subject to search database functionality

At this protocol stage, we have tested and specified search terms (as per Table X1.2 and X1.3 above). As described in the sections above, we also plan to investigate the wider cognitive science concepts identified within the concept map (Appendix 2). We have included a general cognitive science search string above, and hope that this is sufficient to capture wider cognitive science classroom interventions meeting our eligibility criteria. As we describe below, a degree of iteration may be needed and beneficial. Early results and mid-point findings may reveal further cognitive science concepts for which additional searches and search strings (as above) may be required. Any further development of search terms and full search records from all search databases will be recorded and appended to a revised version of this protocol on completion, including the basis for all decisions which tighten or broaden the scope of the review. Once all records located through searches are imported within the EPPI-Reviewer software, all subsequent review methods will be recorded using this specialist software (see below for further details).

The bases for the decision on concept selection and additional searches are set out in the next section on the approach to iteration, below.

Selection of studies

The initial eligibility criteria for the review are set out above. The review will work iteratively though two rounds of activity applying these initial eligibility criteria as a first screen. Rounds of increasingly detailed and rigorous criteria are applied at each stage. We map these activity rounds to the eligibility criteria applied in Table X1.4, below, which we follow with a further detail of each round.

Table X1.4 – Staged application of initial eligibility criteria.

	Round 1 *Screen Titles and Abstracts*	**Round 2** *Screen full reports*
1. **Population:** Children and young people between 3 and 18 years of age in classroom settings	✓	✓
2. **Interventions/Practices of interest**:		
i. Evaluation of a classroom trials and/or intervention	✓	✓
ii. Uses approaches derived from cognitive science relating to the acquisition and retention of knowledge	(✓)[1]	✓
3. **Study design and outcomes:**		
i. Initially we will include all studies reporting empirical evidence of any type or quality about pupil impact, including reviews, from which we will 'mine' for underpinning studies	(✓)[1]	✓
ii. Studies which have any form/quality of counterfactual.	(✓)[1]	✓[2]
iii. We will flag (but exclude) reviews and meta-analyses for reference mining and to inform the underpinning science or practice review strands	✓	✓
iv. We will flag (but exclude) pieces of relevance to the underpinning science or practice review strands	✓	✓
4. **Language:** Include pieces written in English and peer reviewed (for journal articles).	✓[3]	✓[3]
5. **Bodies of Literature:**		
i. Include all peer reviewed journal articles, and reports based on research commissioned by policy makers, charitable or other non-commercial organisations	✓[3]	✓[3]
ii. Exclude conference proceedings, working papers and master's and doctoral dissertations/theses that were published before January 2017.	✓[3]	✓[3]

[1] Assessing this item will be to some extent possible from title and abstract screening, with definite 'no's' being removed. We will assess after round 2 the false-negative rates of records marked for exclusion based only on titles and abstracts and screen on full papers to ensure accurate coding.

[2] As discussed above, a decision will be made about level of stringency for the study design and quality criteria following an initial literature mapping after round 2 (see below).

[3] These final criteria will mostly be applied during database searching but remain as eligibility criteria during screening for any records for which initial information was missing or erroneous.

After initial calibration, training and quality assurance in the use of the eligibility criteria and screening approach within EPPI-Reviewer, the two rounds of screening will be implemented. This initial calibration will involve three researchers all screening around 30 records and comparing the results, repeating if necessary. After initial screening on the title and abstracts, records selected for further review will go through a second round based on the full text. At each stage, 20% of records will be double screened independently by a second researcher. The comparability of this screening will be reported as a measure of inter-rater reliability, with any discrepancies identified, described and

resolved. In the case of disagreement between the two reviewers in the process of abstract screening, a third reviewer will be involved in the selection process. A flag/code will be used to mark studies which need to be reviewed by other team members to reach agreement and confidence in the coding. Studies that do not meet the core criteria yet have potential value for the sub-strands of the review (e.g., qualitative studies of cognitive science applications) will be retained in a separate folder, for potential use in contextualising the quantitative findings.

The screening and final selection process will be documented in a PRISMA chart produced within EPPI-Reviewer. As part of the searching process, the researchers may utilise tools that assist with the identification and extraction of records, but all records will be checked manually.

Approach to Iteration

The systematic review process is iterative in two respects:

- First, in the range of cognitive science concepts considered within the review. All included studies will meet the core eligibility criterion (2ii. approaches derived from cognitive science relating to the acquisition and retention of knowledge); however, how narrowly/broadly this is interpreted, and where and how to broaden it, will depend on emerging evidence.
- Second, in terms of how stringently the additional eligibility criteria (see below) are interpreted for study exclusion, particularly in relation to methodological quality appraisal criteria are in relation to specific cognitive science concepts. Where there is strong evidence in an area, the criterion will be tightened to include only the strongest and most ecologically valid, causal evidence. Where evidence is weak or limited, additional empirical pieces of lower quality will be retained to allow an account of the cognitive science concept and pave the way for future research.

For both of these, the central consideration is the potential for informing practice through gathering and presenting robust, and relevant evidence on cognitive science-informed interventions and practices.

Bases for iterative inclusion – A decision on whether and which additional concepts will strengthen the systematic review and are feasible to include within its scope of the systematic review will be based on the following ***bases***:

a. The ***quantity and quality of evidence*** emerging from data gathering for the five core review concepts. This includes practical considerations, around the resource envelope available to the review, as well as quality considerations around, for example, coverage of related and complementary cognitive science concepts and their operationalisation emerging from the search database after the first round of searching and screening (see below).
b. ***Emerging findings from the underpinning science review*** examining specific cognitive science concepts related to the acquisition and retention of knowledge (see Section 3b for objectives), identifying specific cognitive science concepts, related concepts, terminology and their educational applications.
c. ***Emerging findings from the practice review*** (see Section 3c for objectives) identifying variants in classroom practices, interventions and the terminology surrounding them that derive for cognitive science.
d. ***Advice from the advisory group*** relating to priority areas for investigation.

Criterion for iterative inclusion – A decision on whether and which additional concepts will strengthen the systematic review and are feasible to include within its scope of the systematic review will be based on the following ***criterion***:

1. **Quality** of study in relation to
    a. Internal validity for answering its own question
    b. **Relevance** to our questions
    c. Ecological validity

(See below for details of quality appraisal, including extraction and appraisal tools)

2. **Topic** boundaries
    a) Definition of cognitive science and its boundaries
    b) Application to learning and its boundaries – e.g., teaching and learning in formal learning contexts
    c) Meaningful connection with targeted interventions (and or other well-established interventions)
3. Amount of **detail** regarding factors that affect the potential of studies to inform guidance:
    a) Detail provided regarding teaching and learning processes
    b) Clarity of agency of intervention agent (teacher, machine, learning support assistant)
    c) Detail provided regarding relationship with particular subject or phase of the curriculum
    d) Detail provided regarding CPD provided to teachers
    e) Detail provided regarding links with relevant, broader school policies
4. Any known **conflict of interest**, relating to the independence of evaluation and its design and methods (including outcome measures).
5. Number of studies that can be accommodated within the **resource** envelope

These additional criteria will be applied following the initial general screening (above) on the remaining records. Several items (e.g., relating to internal and ecological validity) have coding items already built into the EEF education database extraction tool (see below). Where standard coding items are not included, additional items will be added. For purposes of screening, these criteria will be coded as closed-response ordinal/binary ratings (e.g., low, medium, high level of detail provided), increasing efficiency (for records ultimately excluded) and enabling more transparent reporting via a PRISMA diagram.

Following this additional coding, the weight of evidence will be assessed using these criteria in a mid-point review ***by cognitive science concept/area***. At this stage selected studies based on the application of the additional eligibility criteria will be progressed to a final round of data extraction, with all decisions fully recorded in EPPI-Reviewer. This will allow for additional more detailed (open-response) and broader set of codes used for extraction than used for screening, while still based on the same criteria.

Duplicates: We will remove all duplicate records. We will in the first instance include multiple publications from the same study or body of work but will subsequently remove any superseded by other related publications associated with the study. Where multiple studies are reported within a single publication, we will apply eligibility criteria to publication sections or chapters pertaining to individual studies and treat eligible sections as single records.

Data extraction and management

Data will be extracted from the selected papers, using a coding framework based on several parts:

1) The EEF main data extraction tool
2) The EEF effect size data extraction tool
3) Our quality assessment tools used in iterative inclusion (above), comprising:
    - The Revised Cochrane Risk of Bias Tool (2)[36]
    - The Cochrane GRADE Tool
    - Items for ecological validity from the EEF extraction tools
    - Additional quality appraisal items for **relevance**, **topic**, **detail** and any **conflict of interest** (as above). NB. There are conflict of interest items in the EEF extraction codesets (e.g. for developer-led evaluations) which we will use.
4) Codes produced from our underpinning cognitive science review and practice review as well as input from the advisory group to flag and categorise studies, and extract data relating to cognitive sciences concepts evident within the interventions.

Our data extraction tools (including EEF, risk of bias, and review-specific tools) are provided in below. We provide further details of quality appraisal below.

As with the earlier screening, initial calibration will take place with three researchers coding around 30 records and comparing data. During the process of data extraction, queries will be flagged on the EPPI-reviewer system, and there will be close coordination of the team to ensure quality control. All members of the team working on data extraction tasks will keep detailed records (wherever possible in EPPI-reviewer) and confer with each other should any problems arise.

At the data extraction stage, 20% of records will be double-coded independently by a second researcher. This applies to all extraction items (such as those relating to quality and effects, described below). The comparability of this coding/extraction will be reported as a measure of inter-rater reliability, with any discrepancies identified, described and resolved. In the case of disagreement between the two reviewers, a third reviewer will be involved in the selection process.

Key details of the entire screening and subsequent data extraction process will be presented in tables and a PRISMA diagram produced in EPPI-Reviewer. These reports will contain, for examples, key information relating to the search (e.g., how many studies were included/excluded for each search, origin of studies (by continent and context), quality ratings per category, and so on).

Appraisal of included studies

Appraisal of included studies

We will quality appraise studies using the criteria set out in the additional criteria, above, and repeated for convenience below:

Quality of study in relation to:

a) Internal validity for answering its own question
b) Relevance to our questions
c) Ecological validity

[36] https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials

These three criteria are operationalised in our coding/extraction framework using code-sets in four areas (again, given above, but repeated here for convenience). Our quality assessment tool used in iterative inclusion (above), comprising:

Quality appraisal tools:

- The Revised Cochrane Risk of Bias (2) Tool[37]
- The Cochrane GRADE Tool[38]
- Items for ecological validity from the EEF extraction tools
- Additional quality appraisal items for **relevance**, **topic**, **detail** and **conflict of interest** (as above)

In overview, we will use 1) the Cochrane Risk of Bias (RoB2) tool to assess internal validity at individual study level; 2) the GRADE assessment criteria to quality assess at the level of cognitive science concept areas[39]; 3) selected items from the EEF tools along with intervention frequency and duration codes to assess ecological validity[40]; and additional items to code for relevance, topic boundaries and detail.

We provide brief further details of each of these below, and the full code sets in this appendix.

Revised Cochrane risk-of-bias tool for randomized trials (RoB 2) (this appendix, below)

Several quality assessment tools were considered for the purpose of assessing risk of bias. These were narrowed done to the Quality Appraisal Checklist for quantitative intervention studies (NICE), the Cochrane Risk of Bias tool (Rob 2), and the Quality Assessment Tool for Quantitative Studies (EPHPP). While all three of these tools provide the means of assessing study quality across several core domains (I.e., selection bias, allocation to groups, outcome measures, reporting bias), a close inspection of available tools indicated that the **RoB 2: A revised Cochrane risk-of-bias tool for randomized trials** was the most applicable and appropriate for the purposes of our systematic review.

RoB 2 is structured into a fixed set of domains of bias, focussing on different aspects of trial design, conduct, and reporting. Within each domain, a series of questions ('signalling questions') aim to elicit information about features of the trial that are relevant to risk of bias. A proposed judgement about the risk of bias arising from each domain is generated by an algorithm, based on answers to the signalling questions. Judgement can be 'Low' or 'High' risk of bias, or can express 'Some concerns'. The RoB 2 will be applied to each of our included studies using EPPI-Reviewer.

Cochrane GRADE Tool (This appendix, below)

To summarise the overall strength of evidence in each cognitive science area for all included studies, we will use the GRADE criteria. We will report the results by GRADE factor in a summary of results table by cognitive science area. We will use our coded data (as above) and follow the GRADE handbook guidance[41] to assess risk of bias, indirectness, inconsistency, imprecision, publication bias and dose effect. This assessment tool has been developed in a medical setting, and even in that setting is under development. However, we believe there are analogous considerations in the educational trial literature which allow its application. For example, we are interpreting indirectness in terms of

[37] https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials
[38] https://training.cochrane.org/grade-approach
[39] To support the coding for the RoB2 and GRADE tools, we will draw on items from the EEF data extraction (and effect size extraction) frameworks relating to study quality (i.e. participant group assignment process and level, design strength for causal inference, sample size, attrition/drop-out, group comparability, outcome measure quality).
[40] We have been advised on suitable items by Prof. Steve Higgins
[41] https://gdt.gradepro.org/app/handbook/handbook.html#h.fzuoa9x107cu

translation and ecological validity and dose effects in terms of frequency and duration of interventions, and are coding for these within our tools (RoB, Ecological validity, and additional quality items). We will use a best-fit judgement and the 4-point GRADE scale when upgrading and downgrading.

Ecological Validity Assessment (This appendix, below)

Given the discussion of ecological validity and translation in the opening sections, as well as the advice from the advisory group around maximising benefits through clear application to particular contexts and settings, we plan to use several EEF extraction and effect size items and create several additional codes which will be used to assess ecological validity (again, reported in an overview table by cognitive science concept area). When assessing external/ecological validity, we will assess whether research results represent what happens in typical classroom teaching and learning in the population of interest (in this case school pupils aged between 3 and 18). Key variables for the assessment relate to how realistic the study was (in relation to typical classroom teaching and learning), the number and type of schools involved, who was responsible for delivering the intervention, and the duration and frequency of intervention sessions (see this Appendix, below).

Additional quality appraisal items for relevance, topic and detail

Additional data extraction codes will be used to assess to assess **relevance** to our questions, relevance in terms of **topic** boundaries, the quality of reporting in terms of **detail**, and any **conflict of interests** (see above). These are simple closed-response codes which a) allow us to make and transparently record quality appraisal decisions and b) flag records for more detailed data extraction in open-response items to support analysis, synthesis and reporting.

Effect size calculation

We will record all effect sizes using the EPPI-Reviewer coding items. As discussed below, we do not expect to have a sufficient number of homogeneous studies to conduct a meta-analysis of the results. We will however extract and report effect sizes for all studies where this is applicable, reporting these using the author's original preferred calculation within the narrative review. Where an effect size is not reported within the original study, but it can be calculated using the data presented, we will use Hedge's g, as recommended by the EEF (see Borenstein *et al*., 2009, p.27 for the formula); with the numerator being the difference in means between the two respective groups, and the denominator being the pooled standard deviation. With a sufficient number of effect sizes (for sufficiently homogenous interventions), we will use the I^2 statistic and estimate τ^2 to consider heterogeneity of the sample and sub-samples of studies (Borenstein *et al*., 2009, pp.114 and 117).

Unit of analysis issues

Again, while a meta-analysis is not planned, we will record the level of randomisation for randomised trials during the data extraction process using the standard EEF data extraction codes. We will also record group sizes and units of analysis when extracting data. We will calculate weighted mean effects using a fixed effects model and provide an overview using a forest plot. We will use the meta-analysis tools within EPPI-Reviewer (and its integration with R and the Metafor package) to calculate the effect sizes and statistics.

Dealing with missing data

Wherever possible, missing values will be calculated from the paper. This can be achieved in instances where effect sizes are not reported, but group scores, sizes and standard deviation statistics are. If

that is not possible, the authors of the papers will be contacted by email and asked to supply the missing data if it is deemed of potential importance to the review findings.

Data synthesis

At the highest level, we will report summary findings for each cognitive science concept area using the GRADE assessment tool, as described above, GRADE ratings will be provided in a summary table which also provides overview information about the concept and the studies within the area and our key findings for the overall concept/intervention area.

Within results sections for each cognitive science concept area, results will be reported by a combination of a study-level summary of results table followed by a structured narrative synthesis approach. The narrative synthesis approach will be structured into groups of studies based on our quality appraisal criteria, as above. Summary judgements from our quality appraisal (i.e., on ecological validity, risk of bias, question and topic relevance, and detail) will be provided in the overview table. The summary table will be based on selected, closed-response fields from the EPPI-Reviewer database in the focus area/report section.

We will follow the summary table in each section by summarising in narrative form, the overall (as per the GRADE rating) and study-level weight and quality of evidence in the area. This makes transparent and communicates to the reader the confidence vested in particular studies and groups of studies as a basis for inferences within the narrative synthesis. We will group studies based on their quality appraisal and proceed to report the narrative synthesis in relation to quality level/groups. For example, we might begin one section by reporting that that there were a group of studies with high internal validity (perhaps based on well-conducted, randomised controlled trials) but low ecological validity; we will indicate which these are in the summary table and then provide a narrative synthesis of the findings of these; then – in this example – perhaps we have a group of medium validity studies but with higher ecological validity and greater detail in intervention process and implementation reporting; we would then (again with transparency about which these studies are and the overview table for readers to consult) discuss how these studies support, refine or refute those of the first group.

Synthesis and reporting will therefore be centred on clear organisation and presentation of data to address the research questions in each strand (as described in this protocol), with clear indication of quality characteristics of the studies within the synthesis throughout. The exact approach to grouping and reporting the synthesis will be decided when all selected papers have been identified for review and in collaboration with the project advisory group. Possible options in terms of structure is to organise the findings around the type of intervention/practice, around different subjects or age groups. If, for example, our advisory group strongly advises reporting of studies in the primary age context separately from the secondary age ranges we would organise extracted data, synthesise and present results by age range. Crucially, we would retain our structured narrative synthesis approach, as described above, of first describing the weight and quality of evidence in the section for synthesis before analysis of studies grouped by quality appraisal results.

At this stage we are not deciding on the minimum numbers of studies required for the analysis, as this will depend on the amount of studies identified in total and the specific approach for organising the data. We are also not planning a meta-analysis of the quantitative data, as the studies are likely to be too heterogeneous and include too many contextual factors. However, we will record the key details required to enable researchers to potentially conduct a meta-analysis in the future, something likely

to be of value once the body of evidence in this area has grown. Any effect sizes (as well as any other main study findings and results) will be reported in overview tables of study details and outcomes.

Investigation of heterogeneity

The present review will ensure that all potentially relevant study characteristics are captured during the data extraction process and considered and reported within the narrative synthesis. To this end, we have developed a carefully constructed data extraction tool, based on and to supplement items on the tool provided by the EEF with added elements derived from our scoping work and advisory group meeting.

Objective 3 of the review relates to differential effects on any groups of pupils, and the data extraction tool thus includes information about the educational stage, age, gender and ethnicity of the children, as well as the proportion eligible for Free School meals, and the proportion who have SEND. These factors others within the EEF extraction template will be coded during the data extraction process and used to evaluate whether there are key areas of difference. Based on the discussion in the first advisory group, we will also record information about the specific subject that the reported intervention was directed towards.

It is possible that additional moderators will be identified throughout the review process (e.g., through the underpinning cognitive science review, the practice review or when reading the full texts of included papers). To ensure that the data extraction tool includes all relevant moderators, we will at an early stage of the study, review it and revise if necessary. The educators and cognitive neuroscientists present in the 1st Advisory Group meeting (Appendix 2) emphasised the importance of age, with concerns that some cognitive science-informed approaches may show greater effectiveness for older children than younger children. The age of participants in each study will be recorded and potentially used as a means of performing subgroup comparisons. We will record the basis for all sub-group/moderator analysis prior to conducting analysis and report all (including null) results of these.

Sensitivity analysis

A meta-analysis is not planned as part of this review. As noted above, relevant data will be captured so that this could be performed at a later date.

Protocol implementation and deviations

The protocol was implemented with high fidelity. As discussed in the limitations section of the main report, above, the largest deviation from the protocol was the limitations in the extent to which sub-group and heterogeneity analysis was possible. The above description therefore serves as a description of the actual as well as the planned method.

In this section we describe deviations from the protocol and details of how approaches described on a more general level in the methods were operationalised. We also report inter-rater reliability ratings from the screening and eligibility assessments. We note that the planned methods allowed for a degree of judgement and iteration in the process within pre-planned bounds. Decisions about grouping studies by strategy and for analysis were made before the analysis was carried out. To account for the inevitable need for judgement and expertise during the analysis (as discussed in the limitations section, above) the analysis was designed to be as transparent as possible to enable any biases or misjudgements to be apparent to readers and readers to be in a position to draw different conclusions.

Searching

The searching went as planned. We have retained all search records, including number of results, RIS files with references and all final search strings. We can provide these records on request. The only slight deviation from the protocol plan with regards to searching was an issue encountered with the JSTOR search database.

We conducted searched in several batches on all databases. The search returning the most results on the JSTRO database[42] returned 8,040 results. This was our general search for cognitive science. There were three other searches for the specific focus techniques/concepts. Our intention was to retrieval all of these results to import into EPPI-Reviewer. However, due to a technical problem with the JSTOR database, only the first 5,050 results (202 pages) from the general search could be exported as a RIS file, after several attempts on different browsers, accounts on different dates, we could not resolve this issue. The results were ordered in terms of relevance and so we conducted an assessment of whether the remaining results were likely to include relevant results. The last 500 of those 5050 results were inspected for relevance. The ver-whelming majority of these did not meet even broad eligibility criteria. With the low relevance of these results, we decided to import into EPPI-Reviewer only the first 5050 results from this one (of four) searches in JSTOR. Therefore, our general searches on JSTOR included 5050/8040 results and the three strategy-targeted JSTOR searches all went ahead as planned. General searches were all completed as planned for all other databases.

Eligibility Assessment for Research Methods and Design

In the protocol we primarily discussed methodological quality in terms of research design. Our focus was on identifying trials. We considered whether non-randomised designs could add value to the review and decided (as per the information above) to retain all studies with any form of counter-factual, including comparisons created by statistical methods. Our intention was to potentially tighten this in cognitive science areas with a sufficient weight of evidence. In the review the need to organise studies along these lines was not often required. The original search terms included terms such as 'trial', 'experiment', 'test' and 'quasi-experiment'. We did not have many studies in the database using more sophisticated quasi-experimental techniques (e.g., difference in difference, propensity score matching) and included and reported the small number we did find where they met the wider criteria. Whether because these terms did not pick them up, or because the lack of studies using these research designs, the need to make fine-graded distinctions by research design was not as prominent as expected.

One methodological aspect that received more focus than originally expected was the sample size and range of study locations/populations. We had not set in advance a specific criterion for eligibility study size (e.g., including studies with pupil N > 100). The size of the study and range of locations is a question of both ecological and internal validity for the studies. In practice it was assessed within our ecological validity assessment (see below) because we had not set a specific rule in advance for this.

Final Eligibility Assessment

The approach to screening and eligibility assessment, as planned and described above, was iterative and multi-stage. The final aspect of this, after the general screen was a quality appraisal to identify

[42] (classroom OR teach* OR school OR lesson) AND (intervention OR trial OR evaluation OR experiment) AND ("cognitive science" OR "learning science")

the most pertinent and high-quality results within the overall database for a) in-depth assessment and b) inclusion in the evidence review. The protocol identifies several quality appraisal tools, as follows:

Quality appraisal tools:

- The Revised Cochrane Risk of Bias (2) Tool[43]
- The Cochrane GRADE Tool[44]
- Items for ecological validity from the EEF extraction tools
- Additional quality appraisal items for relevance, topic, detail and conflict of interest (as above)

At this point in the screening and eligibility assessment process (see Appendix 4 for a PRISMA flow diagram overview) we still had 700 remaining records. It was not feasible to assess Risk of Bias for this number; the GRADE tool was planned for the final assessment of evidence by strategy; and the EEF extraction tool items relevant to ecological validity were a) not finely graded with level descriptors or b) designed to systematically combine to reach an ecological validity assessment.

What was needed was a tool that was able to identify the relevance and potential value of studies within the 700 records, and identify records to exclude. We decided that the ecological validity items and additional quality appraisal items as specified in the protocol were suitable but a) needed refinement and b) needed to be brought together into a specified, coherent tool for this eligibility assessment to be done in a fully systematic and transparent way. The research team brought items together in a 'Final Eligibility Screening Tool' (see the final section in this Appendix). We presented the ecological validity assessment strand of this tool to the advisory group in the second advisory group meeting who advise about how ecological validity and the tool more generally could be used to identify the most relevance and potentially valuable studies. Following the advisory group meeting, this tool was finalised and applied. Further details of this tool and how it was applied are provided in section A3 in the main body of the report, describing the review search process and methods. A key point from this in relation to protocol deviation was the tightening of the cognitive science relevance criteria applied in the first stage of screening. The additional explanation of this has been included in Section A3 and reads as follows:

On the cognitive science relevance criterion: This assessed a) the study's relevance to our cognitive science strategy definitions and focus questions, and b) the strength and clarity of the test of the strategy and/or principle. For this we looked for a clear and relevant counter-factual and controlled conditions. Relevant counterfactuals are strategy-specific: each cognitive science concept implies alternative strategies that are not aligned with the principle in question e.g., massed versus spaced practice, restudy or re-presentation versus retrieval practice, and so on (see definitions in each of the evidence review sections); for purposes of transparency, several studies not meeting this criteria are detailed and indicated in the overview of studies for each strategy, but not included in the results. The requirement to have controlled conditions extended the design criteria (3, above, concerning the need for experimental or quasi-experimental designs) to also require defined interventions/conditions that would test a cognitive science strategy or principle. There were studies, for example, where a cognitive science strategy or principle was an incidental or minor aspect of a study designed to examine another question. The need for this second stage of assessing relevance stemmed in large part from the

[43] https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials
[44] https://training.cochrane.org/grade-approach

challenges of operationalising the concept of 'cognitive science informed' intervention. This concept did not lend itself to pre-specification and needed to be assessed against the actual data.

Concerning the second eligibility assessment of this criterion: Given the size of the database at this point still being unfeasible, we tightened the definitions of cognitive science and ecological validity during this process. The latter was tightened via the use of the ecological validity screening tool (see Appendix 1) which went beyond the population, setting and outcome criteria from the initial screen. The cognitive science relevance was also tightened from the initial screen. As noted above, 'cognitive science informed intervention' was a challenging concept to operationalise. In our first round of screening, we erred on the side of caution, retaining studies with more tenuous links to cognitive science, or vaguer operationalisation and testing of cognitive science strategies. Having initially looser interpretation of the criteria and then tightening allowed us to build up familiarity with the evidence base and the borderline-eligible studies, enabling us to be more confident of consistency when applying the tighter criteria. We described the process and reasons for the need for iterative application of criteria in the original protocol. In effect, the second round of eligibility assessment organised studies into four groups: high, medium and low priority, and exclude – where the latter was the result of tighter relevance criteria and a more precise ecological validity assessment tool. This stage resulted in 201 more exclusions.

All assessments were coded in EPPI-Reviewer and all overall ratings were assessed by two researchers for all papers. In the main results sections, we report results of the use of this tool in terms of high medium and low eligibility across the tool assessment areas. Use of this tool enabled us to systematically identity 43 studies for Risk of Bias assessment and in-depth analysis and the final database of N = 295 studies, down from the 700 studies retained after the general eligibility assessment. In sum, the items within this tool were drawn from those in the original protocol (ecological validity and additional items), but were refined and brought together into a single coherent tool by the research team, in consultation with the advisory group and applied before other quality appraisal tools were implemented.

Inter-rater Reliability of Screening and Coding Rounds

As per the protocol, we double-coded 20% of records at each stage of screening. We double coded the first 20% of all records. Discussion and reconciliation of judgements following this was designed to resolve disagreements for specific records and improve inter-rater reliability on all subsequent items. Two researchers (RL and TP) coded these records and another (CJ) adjudicated any disagreements. During the first screening stage, looking at titles and abstracts for the general eligibility criteria (as above) we screened all of over 40,000 records (see Appendix 4 for overview). 20% of these, 8,151 records, were double-coded. For these, there was a difference in the inclusion/exclusion recommendation for 623 (7.6%) records. Discussion of these revealed that the main disagreements stemmed from studies at the boundaries of relevance for cognitive science informed interventions. It was not clear cut whether a) our focus strategies were adhered to sufficiently tightly and b) whether the paper is explicitly informed by the basic science, as described in the protocol. Many of these border-line cases (whether or not included in the first screen) were eliminated when screening on full text as – where there was not information to make a confident judgement – the record was retained until the full text could be inspected. We judge it to be very unlikely, but not impossible that many high-quality, high-relevance studies will have been excluded.

After screening using the title and abstracts, we moved to a screen using the full text, based on the same criteria (we split the screening in this way solely for purposes of efficiency). At this full-text stage

we started with double-coding of the first 20% of records, using the same approach. Of 454 studies double-coded, there were 73 disagreements on inclusion/exclusion (16.1%). Again, discussion and reconciliation of these revealed that the largest issue was the boundaries of the definitions of cognitive science strategies and making a clear cut between low-relevance and ineligibility. Again, we judge the number of medium or high relevance studies not being included to be very low. We have discussed the conceptual and definitional boundaries of the review at length in the main body of the report in both Part A and B.

The final stage was the eligibility assessment (see Appendix 4 for overview). For this stage, as the review resources made it feasible, the overall rating on the Final Eligibility Assessment Tool (end of the Appendix) for all records (N = 700) were assessed by two members of the team. One researcher (RL) coded all records and then the review principle investigator (TP), after discussion, made the final judgement and recorded all items for which the rating was changed. All changes were across adjacent categories (e.g., high to medium, or low to medium). As with previous sections, the main reason for changes was the difficulties defining the boundary of the focus cognitive science strategies. We also down-graded several studies with non-randomised designs (see above). The mixed strategy programme had a high rate of change. The records were originally coded in terms of the extent to which they adhered to specific strategies, but with mixed strategies (and sometimes wider strategies such as feedback) tested in these studies, we tended to rate this as lower eligibility. We made the decision to include mixed strategies as a group in its own right and this resulted in 4 studies being re-graded more highly.

Table X1.5 – Final Eligibility Assessment Overview

	High	***Medium***	***Low***	***Total studies***	***Rating changed***	***%***
Spaced/Distributed Practice	4	22	19	45	8	17.8%
Interleaving	6	6	4	16	3	18.8%
Retrieval Practice	4	34	26	64	9	14.1%
Working with Schema	4	34	49	87	4	4.6%
Managing Cognitive Load	7	86	59	152	8	5.3%
Cognitive theory of multimedia learning	7	70	45	122	10	8.2%
Embodied Learning and Physical Activity	1	13	12	26	2	7.7%
Mixed Strategy/ Programme	5	3	7	15	4	26.7%

Effect size calculation

In the original protocol, we described that we would use Hedge's g, as recommended by the EEF, where an effect size is not reported within the original study but can be calculated using the data presented. However, many of the studies included frequently reported Cohen's *d* and, therefore, we chose to adopt this effect size throughout the review. Some studies also report measures based on strength of association rather than magnitude of effect (e.g., η2). These have been included in the overview summary tables for each strategy review group.

Data Synthesis

Another aspect where there was a deviation from the review protocol was in relation to synthesis of the evidence. The original hope was that there would be sufficient number of high priority studies with a low risk of bias that we would be able to use as the main basis for our findings. The protocol described our plan to report in a 'structured narrative synthesis' the high studies alone, followed by a short discussion of any additional supporting evidence from the medium priority studies. The number of high studies however was low and not sufficient for many 'stand-alone' judgements due to their eligibility. Our approach was therefore to describe the high priority studies in detail; report the whole group in a summary table with the high studies identified; and, finally, to then summarise and base our results on the overall group (of high and medium priority studies). In other words, the original plan was to report high priority studies and then moderate and extend the results using the medium studies, but in the implementation, we considered all studies in as a group. This introduce the limitation that we were not able to conduct a risk of bias or in-depth data extraction for all of the (medium) studies included in the analysis. We discuss this limitation in the main body of the report in the limitations section (C3). In terms of protocol deviation, the inclusion of medium priority studies in the analysis required us to introduce a greater degree of systematicity in our synthesis approach. We used the GRADE assessment tool to systematically describe results in each section, the risk of bias assessments for the high priority studies and the level of confidence in the findings. The original plan was to use the GRADE tool to provide a benchmark and systematic confidence level for the main results after structured narrative synthesis. In sum, without the originally planned separation of high and medium studies (and removal of studies with moderate or high risk of bias from the former) the GRADE tool was far more central to the process of synthesis than originally intended.

Examination of Population Heterogeneity

The final area in which the implementation of the protocol resulted in a deviation was in relation to examining heterogeneity. We originally planned to conduct detailed analysis of moderating factors in relation to the *findings* of studies. The studies for review however were too numerous, heterogenous and the reporting of findings in the studies did not allow sub-group reporting and analysis at the level of findings. The exception to this was for pupil prior attainment: many studies broke down results by high versus low attainment and given the particular relevance of this to many cognitive science strategy theories, we have included reporting of study findings by prior attainment where this was possible. However, the details of the effect of interventions on other sub-groups was not sufficient to break down findings further. What we did do was explore differences in student population and intervention focus/context in detail in our description of studies – in terms of which subjects, ages groups and countries were represented by the studies in each strategy area. In contrast to the protocol expectations, it was not possible to go into further detail with regards to ethnicity, socio-economic disadvantage, or gender. We were confined to the most consistently reported factors of age; location; number of schools, classes, pupils; and subject/topic areas. As we explain further in the limitations section in the main body (C3), our analysis of variation in the evidence has a) been done in terms of which age groups, study locations and subject areas were represented in the studies rather than in a breakdown of study findings by these populations, and b) was not able to extend to the additional factors of ethnicity, socio-economic disadvantage, or gender as originally hoped.

Timeline

Below is the original planned timeline. Note that due to circumstances around the Covid-19 Pandemic, the project completed several months later than originally planned and we were not able to carry out the school case study visits. We used the resource for these case studies for a questionnaire survey with follow up interviews, and additional practice literature review. See Appendix 13 for further details of the practice review methods.

Activity	M	J	J	A	S	O	N	D	J	F	Lead
Phase 1 - Set-up and project management											
Set-up and Project Management (ongoing)											TP
Scoping work (inc. research and practice)											TP/all
Create and agree review protocols											TP
Phase 2 - Searching											
Database searching, duplicate removal											TP/RFs
Practitioner survey and initial analysis											DY/AF
Searches for Underpinning Evidence Review											KS
Searches for Practice Review											DY/AF
Abstract/title screening using inc./excl. criteria											TP/RFs
20% Double-blind Screening											TP/RFs
Phase 3 - Extraction											
Screening on full text											TP/RFs
Initial quality appraisal, finalise mapping and extraction strategy											TP/RFs
Coding and Data extraction in EPPI Reviewer											TP/RFs
Complete underpinning evidence review											KS
Complete practice review, with survey evidence											DY/AF
20% Double Coding and mid-point analysis											TP/RFs
Data verification											TP/RFs
Phase 4 - Analysis and Synthesis											
Mapping and quantitative summary											TP/RFs
Additional data extraction for pertinent and high-quality studies											TP/RFs
Interrogation of evidence and consultation including translation review											TP/all
School case study visits inc. prep											DY/AF
Phase 5 - Write up and project completion											
Drafting and finalisation of final review report											TP/all
Drafting and finalisation of school-facing publication											TP/all
Archiving and project completion											TP

Eef main data extraction tool

Publication information	Publication Type	Journal article Report Dissertation or thesis Technical report Book or book chapter Conference paper Other
Research type and method	Name of intervention	
	Description of the interventions	
	Objectives of intervention	
	Is there more than one treatment group	Yes No Not specified or N/A
	Assignment of participants	Random, Non-random/Matched Non-random/non-matched prior to treatment Natural sample Retrospective Quasi Experimental Design Regression discontinuity Unclear
	Level of assignment	Individual Class School-cluster School whole site Region Not provided
	How realistic was the study?	High/Low/Unclear Ecological Validity
Location	Study country	
	Additional information	Name of city, region or district Rural/urban/sub-urban No further information
Educational setting	Preschool/Nursery Primary school Middle school Secondary/High school Residential/Boarding School Private/Independent School Home Further education/Junior or Community College Other educational setting Outdoor adventure setting	
Study Sample	Overall number of participants (both intervention and control)	
	Gender	Male Female Mixed

		No Information
	Age	3 – 18
	Proportion of FSM/low SES children in the sample	Add specific indicators of FSM/Low SES
The intervention	Who was responsible for the intervention?	School Charity/NGO University researchers Local Authority Private or commercial company Other
	Was training provided for the delivery team?	Yes No Unclear
	Who was the focus of the intervention?	Students Teachers Teaching assistants Other education Practitioners Non-teaching staff Senior Management Parents Others
	Teaching/intervention approach	Large groups Small groups Pairs One-to-one Students alone
	Was digital technology involved?	Y/N
	Were parent/community volunteers involved?	Y/N
	When was the intervention delivered?	During regular school hours Before/After school Evenings/Weekends Summer holiday period Other Not specified
	Who was responsible for the teaching at the point of delivery?	Research staff Class Teachers Teaching assistants Other school staff External teachers Parents/Carers Lay persons Peers Digital technology Unclear
	Duration of the intervention	
	Frequency of the intervention	
	Length of intervention sessions	
	Are implementation details and/or fidelity details provided?	Quantitative Qualitative

		No details
	Costs of the intervention	Amount Not-specified
Evaluation of the interventions	Who undertook the evaluation?	The developer A different organization paid by developer An organization commissioned independently to evaluate EEF evaluation Unclear/not stated
	Reported primary outcomes	Standardised test Researcher developed test School-developed test National test or examination International tests
	Curriculum subjects tested	Literacy (first language) • Reading comprehension • Decoding/Phonics • Spelling • Reading other • Speaking/listening • Writing Mathematics Science Social studies Arts Languages Other curriculum test
	Other reported outcomes	Yes No
	If yes, which outcomes	Cognitive outcomes measured Other types of student outcomes Other participant outcomes

Eef effect size data extraction tool

Study design	What type of study design is used for the evaluation of impact?	Individual RCT Cluster RCT Multisite RCT Prospective QED Retrospective QED Interrupted time series QED Regression Discontinuity with randomisation Regression Discontinuity - not randomised Regression Continuity - naturally occurring
	Are details of randomisation provided?	Yes Not applicable No/Unclear
Number of schools	What is the number of schools involved in the intervention group(s)?	
	What is the number of schools involved in the control or comparison group?	
	What is the total number of schools involved?	
	Not provided/ unclear / not applicable	
Number of classes involved	What is the total number of classes involved in the intervention group?	
	What is the total number of classes involved in the control or comparison group?	
	What is the total number of classes involved?	
	Not provided/ unclear / not applicable	
Sample description	What is the sample size for the intervention group?	
	What is the sample size for the control group?	
	What is the sample size for the second intervention group?	
	What is the sample size for the third intervention group?	
	Does the study report any group differences at baseline?	Yes No/Unclear
	Is comparability taken into account in the analysis?	Yes No Unclear or details not provided
	Is attrition or drop out reported?	Yes No Unclear (please add notes)
	What is the attrition in the treatment group?	

	Are the variables used for comparability reported?	Yes No N/A
	If yes, which variables are used for comparability?	Educational attainment • Gender • Socio-economic status • Special educational needs • Other (please specify)
	What is the total or overall percentage attrition?	
	Is clustering accounted for in the analysis?	Yes No Unclear
Outcome details	Are descriptive statistics reported for the primary outcome?	Yes/No
	If yes, please add for the intervention* group	Number (n) Pre-test mean Pre-test standard deviation Post-test mean Post test standard deviation Gain score mean (if reported) Gain score standard deviation (if reported) Any other information?
	If yes please add for the control group	(as previous)
	If yes, please add for a second intervention* group (if needed)	(as previous)
	If needed, please add for the second control group	(as previous)
	If yes, please add for a third intervention* group (if needed)	(as previous)
	If needed please add for a third control group	(as previous)
	Is there follow up data?	Yes No
	Primary outcome	
	Secondary outcome(s)	
	SES/FSM outcome	
Outcome classification	Sample (select one from this group)	Sample: All Sample: Exceptional Sample: High achievers Sample: Average Sample: Low achievers
	Test type (select one from this group)	Test type: Standardised test Test type: Researcher developed test

		Test type: National test Test type: School-developed test Test type: International tests
Effect size calculation (select one from this group)	What kind of effect size is being reported for this outcome?	Post-test unadjusted Post-test adjusted for baseline attainment Post-test adjusted for baseline attainment AND clustering Pre-post gain
Toolkit strand(s)	Arts participation / Aspiration interventions / Behaviour interventions / Block scheduling / Built environment / Collaborative learning / Digital technology / Early years intervention / Extending school time / Feedback / Homework / Individualised instruction / Learning styles / Mastery learning / Metacognition and self-regulation / Mentoring / One to one tuition / Oral language interventions / Outdoor adventure learning / Parental engagement / Peer Tutoring / Performance pay / Phonics / Reading comprehension strategies / Reducing class size / Repeating a year / School uniform / Setting or streaming / Small Group Tuition / Social and emotional learning / Sports participation / Summer schools / Teaching assistants	

Revised cochrane risk-of-bias tool for randomized trials (rob 2)

The following tool will be applied to all randomised trials. We will follow the guidance provided in the handbook when implementing the tool:

https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool/current-version-of-rob-2

Preliminary	Study design	Individually-randomized parallel-group trial Cluster-randomized parallel-group trial Individually randomized cross-over (or other matched) trial
	Intervention definition, Experimental:	
	Intervention definition, Comparator:	
	Specify which outcome is being assessed for risk of bias	
	Specify the numerical result being assessed. In case of multiple alternative analyses being presented, specify the numeric result (e.g. RR = 1.52 (95% CI 0.83 to 2.77) and/or a reference (e.g. to a table, figure or paragraph) that uniquely defines the result being assessed.	
	Is the review team's aim for this result…?	• to assess the effect of assignment to intervention (the 'intention-to-treat' effect) • to assess the effect of adhering to intervention (the 'per-protocol' effect)
	If the aim is to assess the effect of adhering to intervention, select the deviations from intended intervention that should be addressed (at least one must be checked):	• occurrence of non-protocol interventions • failures in implementing the intervention that could have affected the outcome • non-adherence to their assigned intervention by trial participants
	Which of the following sources were obtained to help inform the risk-of-bias assessment? (tick as many as apply)	Journal article(s) Trial protocol Statistical analysis plan (SAP) Non-commercial trial registry record (e.g. ClinicalTrials.gov record) Company-owned trial registry record (e.g. GSK Clinical Study Register record) "Grey literature" (e.g. unpublished thesis) Conference abstract(s) about the trial Regulatory document (e.g. Clinical Study Report, Drug Approval Package) Research ethics application Grant database summary (e.g. NIH RePORTER or Research Councils UK Gateway to Research) Personal communication with trialist Personal communication with the sponsor

Domain 1: Risk of bias arising from the randomization process	1.1 Was the allocation sequence random?	Y/PY/PN/N/NI
	1.2 Was the allocation sequence concealed until participants were enrolled and assigned to interventions?	Y/PY/PN/N/NI
	1.3 Did baseline differences between intervention groups suggest a problem with the randomization process?	Y/PY/PN/N/NI
	Risk-of-bias judgement	Low / High / Some concerns (Calculated using algorithm based on previous items)
	Optional: What is the predicted direction of bias arising from the randomization process?	NA / Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable
Domain 2: Risk of bias due to deviations from the intended interventions (effect of assignment to intervention)	2.1. Were participants aware of their assigned intervention during the trial?	Y/PY/PN/N/NI
	2.2. Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?	Y/PY/PN/N/NI
	2.3. If Y/PY/NI to 2.1 or 2.2: Were there deviations from the intended intervention that arose because of the trial context?	Y/PY/PN/N/NI
	2.4 If Y/PY to 2.3: Were these deviations likely to have affected the outcome?	Y/PY/PN/N/NI
	2.5. If Y/PY/NI to 2.4: Were these deviations from intended intervention balanced between groups?	Y/PY/PN/N/NI
	2.6 Was an appropriate analysis used to estimate the effect of assignment to intervention?	Y/PY/PN/N/NI
	2.7 If N/PN/NI to 2.6: Was there potential for a substantial impact (on the result) of the failure to analyse participants in the group to which they were randomized?	Y/PY/PN/N/NI
	Risk-of-bias judgement	Low / High / Some concerns (using algorithm)
	Optional: What is the predicted direction of bias arising from the randomization process?	NA / Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable
Domain 2: Risk of bias due to deviations from the intended interventions (effect of adhering to intervention)	2.1. Were participants aware of their assigned intervention during the trial?	Y/PY/PN/N/NI
	2.2. Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?	Y/PY/PN/N/NI
	2.3. [If applicable:] If Y/PY/NI to 2.1 or 2.2: Were important non-protocol interventions balanced across intervention groups?	Y/PY/PN/N/NI
	2.4. [If applicable:] Were there failures in implementing the intervention that could have affected the outcome?	Y/PY/PN/N/NI

	2.5. [If applicable:] Was there non-adherence to the assigned intervention regimen that could have affected participants' outcomes?	Y/PY/PN/N/NI
	2.6. If N/PN/NI to 2.3, or Y/PY/NI to 2.4 or 2.5: Was an appropriate analysis used to estimate the effect of adhering to the intervention?	Y/PY/PN/N/NI
	Risk-of-bias judgement	Low / High / Some concerns (using algorithm)
	Optional: What is the predicted direction of bias due to deviations from intended interventions?	NA / Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable
Domain 3: Risk of bias due to missing outcome data	3.1 Were data for this outcome available for all, or nearly all, participants randomized?	Y/PY/PN/N/NI
	3.2 If N/PN/NI to 3.1: Is there evidence that the result was not biased by missing outcome data?	Y/PY/PN/N/NI
	3.3 If N/PN to 3.2: Could missingness in the outcome depend on its true value?	Y/PY/PN/N/NI
	3.4 If Y/PY/NI to 3.3: Is it likely that missingness in the outcome depended on its true value?	Y/PY/PN/N/NI
	Risk-of-bias judgement	Y/PY/PN/N/NI
	Optional: What is the predicted direction of bias due to missing outcome data?	Y/PY/PN/N/NI
Domain 4: Risk of bias in measurement of the outcome	4.1 Was the method of measuring the outcome inappropriate?	Y/PY/PN/N/NI
	4.2 Could measurement or ascertainment of the outcome have differed between intervention groups?	Y/PY/PN/N/NI
	4.3 If N/PN/NI to 4.1 and 4.2: Were outcome assessors aware of the intervention received by study participants?	Y/PY/PN/N/NI
	4.4 If Y/PY/NI to 4.3: Could assessment of the outcome have been influenced by knowledge of intervention received?	Y/PY/PN/N/NI
	4.5 If Y/PY/NI to 4.4: Is it likely that assessment of the outcome was influenced by knowledge of intervention received?	Y/PY/PN/N/NI
	Risk-of-bias judgement	Low / High / Some concerns (using algorithm)
	Optional: What is the predicted direction of bias in measurement of the outcome?	NA / Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable
Domain 5: Risk of bias in selection of the reported result	5.1 Were the data that produced this result analysed in accordance with a pre-specified analysis plan that was finalized before unblinded outcome data were available for analysis?	Y/PY/PN/N/NI
	Is the numerical result being assessed likely to have been selected, on the basis of the results, from...	Y/PY/PN/N/NI

	5.2. ... multiple eligible outcome measurements (e.g. scales, definitions, time points) within the outcome domain?		Y/PY/PN/N/NI
	5.3 ... multiple eligible analyses of the data?		Y/PY/PN/N/NI
	Risk-of-bias judgement		Low / High / Some concerns (using algorithm)
	Optional: What is the predicted direction of bias due to selection of the reported result?		NA / Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable
Overall risk of bias		Low / High / Some concerns	
		Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable / NA	
Where:	Low risk of bias	The study is judged to be at low risk of bias for all domains for this result	
	Some concerns	The study is judged to raise some concerns in at least one domain for this result, but not to be at high risk of bias for any domain.	
	High risk of bias	The study is judged to be at high risk of bias in at least one domain for this result. Or The study is judged to have some concerns for multiple domains in a way that substantially lowers confidence in the result	

Cochrane grade tool

The GRADE tool will be used to make an overall assessment of evidence in a cognitive science concept area. We will adhere to guidelines for the GRADE handbook: https://gdt.gradepro.org/app/handbook/handbook.html

Quality of Evidence Levels Definitions

- High: Very confident that the true effect lies close to the estimate of effect
- Moderate: Moderate confidence that the true effect lies close to the estimate of effect
- Low: Limited confidence that the true effect lies close to the estimate of effect
- Very Low: Very little confidence in the estimate of effect

GRADE Quality of Evidence Assessment Process and Rating

<table>
<tr><th>Study Design <u>At Entry</u> Into GRADE System</th><th>Quality of Evidence on Entry</th><th>Lower Category If…</th><th>Higher Category If…</th><th colspan="2"><u>Final</u> Quality of Evidence Rating (Select One)</th></tr>
<tr><td>RCT</td><td>HIGH</td><td rowspan="4">Risk of Bias
-1 Serious
-2 Very Serious
Inconsistency
-1 Serious
-2 Very Serious
Indirectness
-1 Serious
-2 Very Serious
Imprecision
-1 Serious
-2 Very Serious
Publication Bias
-1 Serious
-2 Very Serious</td><td rowspan="4">Effect Size
+1 Large
+2 Very Large
Dose Response
+1
All plausible confounders would reduce a demonstrated effect
+1
All plausible confounders would suggest a spurious effect when the results show no effect
+1</td><td>HIGH
++++</td><td></td></tr>
<tr><td rowspan="2">Observational Study</td><td rowspan="2">LOW</td><td>MODERATE
+++0</td><td></td></tr>
<tr><td>LOW
++00</td><td></td></tr>
<tr><td colspan="2"></td><td>VERY LOW
+000</td><td></td></tr>
</table>

Final eligibility screening tool

<table>
<tr><td>Adapted from EEF extraction tool</td><td>Ecological Validity</td><td>High/ Medium/ Low</td></tr>
<tr><td colspan="3">Using the following tool (best fit)
<table>
<tr><th>Criterion</th><th>Low Ecological Validity</th><th>Medium Ecological Validity</th><th>High Ecological Validity</th></tr>
<tr><td>System relevance (congruence with English education)</td><td>Jurisdictions with marked cultural or organisational differences (e.g. developing countries)</td><td>Jurisdictions with small cultural or organisational differences</td><td>Highly similar jurisdiction. E.g. Other UK nations, New Zealand, Australia, Canada, USA, Most of Europe (including Scandinavia)</td></tr>
<tr><td>Research context (where the research was conducted)</td><td>In a University and/or laboratory or contrived/artificial setting within a school.</td><td>Some level of artificiality relating to the design or delivery of the intervention. (e.g. delivery supported by researchers)</td><td>No artificiality and a separation between design and delivery personnel.</td></tr>
<tr><td>Participant cohort size</td><td><25 students (per comparison group)/1-2 teachers/single setting</td><td>25-100 students per comparison group/3-9 teachers/3 or more settings</td><td>100+ students per comparison group/10 or more teachers/8 or more schools</td></tr>
<tr><td>Learning outcome</td><td>Abstract/simple content with limited curricular applicability</td><td>Learning outcome with curriculum relevance</td><td>Learning outcome(s) with high/cross-curriculum relevance</td></tr>
</table></td></tr>
<tr><td>Tighten from last round</td><td>Test of Cog Sci technique – considering a) strength and clarity of test/counterfactual/intervention/conditions and b) the boundaries of cog sci definition and focus</td><td>High/ Medium/ Low</td></tr>
<tr><td>Item following consultation with advisory group</td><td>Added value to breadth of evidence for ecological validity factors related to: age, subject, disadvantage, type/aim of learning activity/instructional relation, training/resource requirement.</td><td>High/ Medium/ Low</td></tr>
<tr><td>For screening</td><td>Recommendation
High = (additionally to below) for inclusion for coding in EEF Database
Medium = for inclusion for systematic review
Low (background) = retain for non-systematic review in background/wider report sections.
Exclude = does not meet tighter criteria</td><td>High/ Medium/ Low/Exclude</td></tr>
</table>

Appendix 2: Protocol and Scoping

Advisory group – summary of opening meeting – 10/7/2020

Area 1 – application/translation

The advisory group generally considered the application of cognitive science in the classroom as growing, but nevertheless still in its beginning. There was a sense that cognitive science inspired practices were gaining momentum and furthermore an anticipation that they will be increasingly incorporated in policy.

Approaches inspired by cognitive science take time to implement and embed, as they may differ from 'normal' ways of working. Key barriers to their application were considered to include teacher resistance to change, and questions of how to translate and implement cognitive science concepts and theories in classroom practice. One issue, which was mentioned in relation to this was the danger of separating small or selected aspects of the learning process from a broader conceptual framework and losing the broader dynamics when translating a defined result from cognitive scientific research into the curriculum. Furthermore, the question of how to understand learning was raised, and it was noted that evidence of memorised information is not necessarily evidence of learning in itself.

It was generally agreed that in order for cognitive science inspired approaches to be successful and work well, their particular relevance needs to be clear to teachers. Without a deeper understanding of why a particular approach is adopted, and what its specific benefits are for students, there is a danger of approaches being seen merely as "techniques." Although there is a lot of engagement in schools with issues about cognition, it is questionable whether there is a coherent and consistent understanding across educational practitioners and leaders about what the evidence says and how to make use of it. It is furthermore important to acknowledge the many contextual factors around education and how research evidence can be translated into pedagogy in a system.

Area 2 – more/less promising approaches

It was noted that many of the techniques mentioned in the conceptual map of the project are not new (e.g. retrieval practice) and may furthermore have been practised previously by "hunch" (e.g. quizzes) rather than as officially inspired by cognitive science informed practices. Many of these practices resonate with established practice and good pedagogy, but cognitive science may help us move towards a shared understanding and common language. At the same time, insights from cognitive science may help establish redundant practices that are commonly used, and show why some work and some don't.

When discussing the project conceptual map more specifically, the advisory group commented on the following specific points:

- The first five strategies mentioned in the concept map (spaced practice, interleaving, retrieval practice, dual coding and strategies to manage cognitive load) were seen as being pushed a lot in some schools. It was mentioned that it would be good to look specifically at the classroom-based evidence behind them, as they may not translate from lab to classroom as well as we think. Furthermore, it was mentioned that we need to be clear about what these strategies mean, how we define them, and what they look like when they are done well.

- ‘Brain training’ may be too broad a concept which needs to be defined or narrowed, as it can potentially covers many cognitive interventions. One way of making this more specific could be to separate computerised training.
- ‘Brain-to-brain synchrony’ was mentioned as an area which should perhaps have low priority or which could be included in a more general category on the influence of social interaction and peers on learning.
- Mindfulness (included in stress/anxiety on table) is a promising area relating to effects on cognition, and could be included as an additional strategy.
- Analogies, future basing, active promotion of explicit links, mapping and visual schema as an aid to learning were also suggested as additional areas to potentially add to the map.
- Play was also suggested as a potential area to look at, particularly as younger year groups are included in the review.
- It was suggested that we might need to explore the role of the brain in affective domains and how it pertains to topics like values or character-based education and other more holistic questions.
- While the concept map needs to recognise limits and boundaries to what can sensibly be interrogated, at the same time these boundaries should not be treated as wholly exclusionary, as something being practically challenging to explore doesn’t automatically equate to it not being relevant. Achieving a linkage between lesson objective, method and underpinning science about cognition would be a massive achievement, even if the linkages are somewhat loose.

In terms of contextual factors, age was mentioned as a very important element. Much existing work has been focused on secondary schools and universities, but practices tried and tested in older pupils may not be relevant/effective for younger pupils. SEND and cultural factors were also considered as important contextual factors, but it was noted that there might not be enough studies to make firm conclusions.

Potential project benefits (whole group closing discussion)

The group considered that a successful outcome of the project would be to produce a trusted source of documents that practitioners can refer and relate to. Being able to achieve something that even broadly fits that description would be a major and powerful return from the project. It was considered important that the interpretations of broader findings were linked to particular subjects or phases of schooling. Even in the case of findings being consistent across different phases or subject areas, teachers may still be inclined only to engage with the parts of findings that obviously and overtly relate to their particular subject discipline and/or the phase of education they engage with. The project would help teachers feel more confident about their own professionalism and capacity to interrogate research around cognitive knowledge and research.

Finally, from a policy perspective, success was seen to involve less rigidity and certainty in the expectations about how policies should be translated. This relates to a general and cross-cutting point arising from the meeting, which was to avoid a reductive view, not only of cognitive science, but also of the concept of learning and consequently, of links between the two.

Bibliography for scoping

Books

Agarwal, P. K., & Bain, P. M. (2019). Powerful teaching: Unleash the science of learning. John Wiley & Sons.

Brown, P. C., Roediger III, H. L., & McDaniel, M. A. (2014). Make it stick. Harvard University Press.

Busch, B., & Watson, E. (2019). The Science of Learning: 77 Studies that Every Teacher Needs to Know. Routledge.

Chartered College of Teaching (2020) Cognition and Learning: Applying key insights from cognitive science and psychology in the classroom. *Impact* (8, Spring 2020)

Dehaene, S. (2020). How We Learn: The New Science of Education and the Brain. Penguin UK.

Didau, D., & Rose, N. (2016). What Every Teacher Needs to Know about... Psychology. John Catt Educational Limited.

Harrington, J., Beale, J., Fancourt, A., & Lutz, C. (2020). The 'BrainCanDo' Handbook of Teaching and Learning: Practical Strategies to Bring Psychology and Neuroscience into the Classroom.

Jensen, E., & McConchie, L. (2020). Brain-Based Learning: Teaching the Way Students Really Learn. Corwin.

Kirschner, P. A., & Hendrick, C. (2020). How Learning Happens: Seminal Works in Educational Psychology and What They Mean in Practice. Routledge.

Kitchen, W. H. (2017). Philosophical Reflections on Neuroscience and Education. Bloomsbury Publishing.

Tibke, J. (2019). Why the Brain Matters: A Teacher Explores Neuroscience. SAGE Publications Limited.

Weinstein, Y., Sumeracki, M., & Caviglioli, O. (2018). Understanding how we learn: A visual guide. Routledge.

Reviews, meta-analyses and practice-focused reports

Centre for Education, Statistics and Evaluation (2018) Cognitive Load Theory in Practice: Examples for the classroom. Available at: https://www.cese.nsw.gov.au//images/stories/PDF/Cognitive_load_theory_practice_guide_AA.pdf

Churches, R., Dommett, E. J., Devonshire, I. M., Hall, R., Higgins, S., & Korin, A. (2020). Translating Laboratory Evidence into Classroom Practice with Teacher-Led Randomized Controlled Trials—A Perspective and Meta-Analysis. Mind, Brain, and Education. https://doi.org/10.1111/mbe.12243

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980–2016. Educational Research, 60, 276–291. https://doi .org/10.1080/00131881.2018.1493353

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2020). Implications for educational practice of the science of learning and development. Applied Developmental Science, 24(2), 97-140.

Deans for Impact. (2015). *The Science of Learning.* Deans for Impact. Available: http://www.deansforimpact.org/wp-content/uploads/2016/12/The_Science_of_Learning.pdf

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public Interest, 14(1), 4-58.

Howard-Jones, P. (2014). Neuroscience and education: A review of educational interventions and approaches informed by neuroscience. Education Endowment Foundation, Millbank, UK.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). Organizing Instruction and Study to Improve Student Learning. IES Practice Guide. NCER 2007-2004. National Center for Education Research.

Rosenshine, B. (2012). Principles of Instruction: Research-Based Strategies That All Teachers Should Know. American educator, 36(1), 12.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.

Wittwer, J., & Renkl, A. (2010). How effective are instructional explanations in example-based learning? A meta-analytic review. Educational Psychology Review, 22(4), 393-409.

Weinstein, Y., Madan, C. R., & Sumeracki, M. A. (2018). Teaching the science of learning. Cognitive Research: Principles and Implications, 3(1), 2.

Other papers and policy documents

Aronsson, L. (2020). Reconsidering the concept of difference: A proposal to connect education and neuroscience in new ways. Policy Futures in Education, 18(2), 275-293.

Ofsted (2019) *Education inspection framework: overview of research.* Ref. 180045, Available: https://www.gov.uk/government/publications/education-inspection-framework-overview-of-research

Goswami, U. (2006). Neuroscience and Education: From Research to Practice? *National Review of Neuroscience, 7*(5), 406-411

GOV.UK (2019) *Early Career Framework*. DFE-00015-2019. Available: https://www.gov.uk/government/publications/supporting-early-career-teachers.

Howard-Jones, P. A. (2014). Neuroscience and education: myths and messages. *Nature Reviews Neuroscience, 15*, 817-824. doi:10.1038/nrn3817

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. Learning and instruction, 4(4), 295-312.

Cognitive science concept map (version produced during protocol development)

Cognitive Science Concept Map				
1. Approaches Informing the Design of Classroom Teaching and Learning (Core Focus)				
Technique/Theory from Cognitive Science	*Classroom Teaching and Learning Theorie(s)*	*Synonyms and Conceptual Connections*	*Examples of Classroom Teaching and Learning Practice(s)*	*Concept-Specific Contextual and moderating factors*
Spaced practice	**Distributing learning and retrieval opportunities** over a longer period of time leads to better retention on delayed tests, compared to massed practice	*Synonyms/related terms:* Spacing; spaced repetition; distributed practice; often compared to 'cramming' (opposite concept) *May overlap with:* - Interleaving - Retrieval practice	- Introducing rapid quizzes - Curriculum planning/sequencing	- Delay to the final test matters - Spacing matters (a few days after learning?) - Taught, homework/practice, test (the space can be between different stages of learning, from intro to practice)
Interleaving	**Switching** between different types of problem or different ideas within the same study session leads to better retention on delayed tests, compared to block practice (e.g. abcbcacab instead of aaabbbccc)	*Synonyms/related terms:* interleaved practice; interweaving *May overlap with:* - Spaced practice	- Introducing rapid quizzes - Lesson planning - Varied practice	- Difficult to disentangle from spacing outside of the lab - Could effects be because of spacing rather than interleaving (i.e. is interleaving merely a form of spacing?) - How related should interleaved material be?
Retrieval practice	**Recalling** information from memory can promote long-term learning	*Synonyms/related terms:* test-enhanced learning; testing effect; recall; quiz; mind-maps *Conceptual links:* - Includes elaboration: very broad (integrating and organising new information with what we already know). Involves asking and explaining how things work - Interleaving	- Introducing rapid quizzes - Practice tests - Flashcards - Mind-maps from memory - Blank sheet of paper ('brain dumps'/open recall) - Cloze/Retrieval guides/scaffolding - Elaborative interrogation - Multiple choice tests. Filling in the blanks	- Better after a delay - Short answer might be better than multiple choice, unless this involves retrieval (e.g. plausible distractors) - Needs to be successful and address misunderstandings - May be more beneficial where students have lower working memory - Elaboration-based retrieval is most helpful when students have more background knowledge

		- Spacing		
Dual coding	Introducing concepts using both **verbal** (i.e. words) and **non-verbal** information (i.e. pictures) is thought to increase the chance of remembering that concept, compared to if the stimulus was only coded one way (i.e. words *or* images)	*Synonyms/related terms:* mental representation; multimedia design; verbal/non-verbal information; opposing theory = propositional theory, which claims that mental representations are stored as propositions rather than as images *Links with:* - Cognitive load	- Effective use of text and graphics on slides - Use of diagrams or cartoons/comic strips - Timelines with images - Diagrams - Infographics - 5 minute lesson plan (i.e. visual guide to lesson)	- Controversial theory - Which pictures/words? - How conceptually relevant?
Strategies to manage cognitive load	Attention and working memory, are essential to learning but are '**limited capacity' resources** which can be over-loaded. Cognitive load theory focuses teachers on the efficiency of their explanation and presentation of new knowledge	*Synonyms/related terms:* 3 types (intrinsic, extraneous, and germane); situational demands Links with: - Dual coding	- Worked examples - Segmented presentation - Pre-training - Narration - Strip back extraneous detail - Less 'redundant' information (i.e. not reading words off slides) - Limit background distractions - Optimum classroom seating arrangements	- Expertise reversal effect (i.e. if the instruction fails to provide guidance, low-knowledge learners often resort to inefficient problem-solving strategies, increasing cognitive load) - Gradually fading out guidance best - Situational interest is important - Potential gender-specific effects for different types of cognitive load - High cognitive load sometimes useful?
Concrete examples	Using **specific examples** to understand **abstract ideas**	*Synonyms/related terms:* analogies; models; modelling; real-life examples Links with: - Dual coding	- Analogies - Models/modelling, especially in science (i.e. to help with visualising microscopic/subatomic structures) - Check background knowledge - Using real-life situations/scenarios when explaining new concepts - Students generate their own concrete example	- Number of examples needed? - Can backfire when a) distracts attention, b) surface features are too salient - Important to use varied examples, so that students do not associate meaning of the concept with only one specific example

Brain training	**Cognitive training** programmes that are designed to boost cognitive functioning, either as a whole, or a specific aspect (e.g. working memory)	*Synonyms/related terms:* cognitive training programmes; brain training games; cognitive training software; executive functioning training Potentially, this category could overlap significantly with other concepts, including retrieval practice, game-based learning, spacing, exercise	- Many commercialised programmes (e.g. ACTIVATE by C8, BrainWare) - Can include a number of different tasks/activities, including memory games, puzzles, multi-tasking games	- Can be classroom-based or online (e.g. apps) and/or at home - Can be broad (e.g. for whole class) or specific (e.g. for students with ADHD or learning disabilities) - How transferable/generalisable are the 'skills' developed in these programmes to everyday learning? (i.e. transfer effects) - Age is important
2. Approaches Involving Physical Factors (Wider Concepts)				
Technique/Theory from Cognitive Science	*Classroom Teaching and Learning Theorie(s)*	*Synonyms and Conceptual Connections*	*Examples of Classroom Teaching and Learning Practice(s)*	*Concept-Specific Contextual and Moderating Factors*
Exercise	It is thought that **exercise** increases efficiency of neural networks that are important for learning, whereby episodes and regimes of exercise can improve cognitive function and memory	*Synonyms/related terms:* physical activity (vs sedentary activity); embodied cognition *Links with:* - Interleaving (if used within lesson) - Stress reduction	- Daily mile - 'Educational Kinesiology' programmes, such as Brain Gym -Physical activity can include curriculum content (e.g. recalling times-tables while performing activity)	- Variable format (e.g. additional slot on timetable, start of school day, embedded at start of/within lesson) - Effects can be measured acutely (e.g. working memory immediately after short burst of physical activity) or long-term (e.g. change in cognitive performance following implementation of physical activity regime)
Nutrition/Hydration	**Hunger and malnutrition** can affect cognitive performance. Hunger affects many aspects of cognition (e.g. working memory, attention), as well as emotional factors (e.g. motivation, engagement) Habitual ingestion of caffeine/dehydration can also reduce cognitive function	Interventions may also be implemented alongside other approaches in this category (e.g. sleep, exercise) as part of a broader health programme	- Breakfast clubs - 'Breakfast after the Bell' programmes - Education about nutrition in lessons - Nutrition programmes	- Could be classroom-based, or school-wide approach
Sleep	Sleep is important for rest and for **consolidating** the day's learning in **long-term** memory	Interventions may also be implemented alongside other approaches in this category (e.g.	- Changes to start/finish times of school day	

		nutrition, exercise) as part of a broader health programme	- Sleep education (i.e. promoting knowledge about sleep)	
3. Approaches Involving the Motivational or Emotional State of the Learner (Wider Concepts)				
Technique/Theory from Cognitive Science	*Classroom Teaching and Learning Theorie(s)*	*Synonyms and Conceptual Connections*	*Examples of Classroom Teaching and Learning Practice(s)*	*Concept-Specific Contextual and Moderating Factors*
Mindfulness	Involves **attending** to and **focusing** non-judgementally on whatever is happening in any given moment, including thoughts, feelings, bodily sensations, and surrounding environment. This is thought aid cognitive processing (e.g. attention)	*Synonyms/related terms:* gratitude; relaxation; mindfulness practice; mindfulness meditation *Links:* Could potentially overlap with stress reduction/SEL	- Mindfulness-based stress reduction programmes - Journalling - Affirmations - Mindful walks - Mindfulness breathing exercises - Often implemented as part of broader intervention (e.g., wellbeing/mental health/ healthy minds)	- Can be digital (e.g. app) or in the classroom - Can be school-wide or classroom-based
Stress/anxiety reduction	**Acute or chronic stress** and anxiety can have detrimental effects on higher order cognitive functions, such as working memory	*Synonyms/related terms:* relaxation; breathing exercises; meditation; yoga; wellbeing; anxiety Could potentially overlap with mindfulness/SEL/exercise	- Breathing exercises - Guided meditation - Yoga or other physical activity - Often implemented as part of broader intervention (e.g. wellbeing/mental health/healthy minds)	- Can be digital (e.g. app) or in the classroom - Can be school-wide or classroom-based
Social and emotional learning	Strategies that help students to effectively apply the **knowledge, attitudes, and skills** necessary to **understand and manage emotions**, set and achieve positive goals, feel and show empathy for others, establish and maintain positive relationships, and make responsible decisions. Being emotionally competent can have benefits for academic performance	*Synonyms/related terms:* Emotional intelligence (training); emotional competence; emotion recognition; emotion understanding; emotion regulation/management; resilience; SEAL *Potential overlap with:* - Stress reduction	- Several commercial programmes (e.g. RULER, - Mood meter (to help students learn emotion recognition) - Meta-moment (helps challenge impulses and negative behaviours) - School charters (to establish supportive and productive learning environment)	- Can be school-wide or classroom-based - Can have a compensatory effect (e.g. emotional intelligence can provide extra resources for a student to draw upon if cognitive ability is low) - Emotional intelligence can be conceptualised using a trait (i.e. emotional self-efficacy) or ability (i.e. emotion-related cognitive skills, such as emotion recognition)

		- Mindfulness - Game-based learning	- Games to help students learn emotion words (and thus increase emotion knowledge)	
Reward/game-based learning	Games provide rapid schedules of uncertain **reward/reinforcement** that stimulate the brain's reward system, which can positively influence the rate at which we learn	*Synonyms/related terms:* educational games; gameplay; positive reinforcement; operant conditioning; reward *Potential links:* - Concrete examples - Retrieval practice	- Minecraft/Lego can be used to help with visuospatial learning - Competitions - Reward/behaviour reinforcement charts	- Enjoyment of game is an important factor - Can involve creativity (e.g. Lego/Minecraft building)
4. Approaches Involving Direct Measurement or Manipulation of Neural Activity (Wider Concepts)				
Technique/Theory from Cognitive Science	*Classroom Teaching and Learning Theorie(s)*	*Synonyms and Conceptual Connections*	*Examples of Classroom Teaching and Learning Practice(s)*	*Concept-Specific Contextual and Moderating Factors*
Transcranial electrical stimulation	Applying small **currents** to the scalp can benefit some cognitive functions and learning processes, potentially by increasing neuroplasticity	*Synonyms/related terms:* TDCS; TACS; TRNS; evoked potentials	Not applicable - not yet established as classroom practice	- Transcranial electrical stimulation may improve learning difficulties in atypical brain development?
Brain-to-brain synchrony	Brain-to-brain synchrony - the coupling of behavioural and biological signals during social contact - is a possible **neural marker for dynamic social interactions**, likely driven by shared attention mechanisms. Greater synchrony may lead to more positive learning outcomes	*Synonyms/related terms:* brain coordination; EEG; fMRI; BOLD; neural synchrony; interbrain synchrony; social connectedness; dyadic interactions; group dynamics; student engagement	Not applicable – not yet established as classroom practice	- Social context is important - Potential for investigating teacher-student dynamics and student engagement - Can depend on how much people like each other (close relationships = greater synchronisation)

Appendix 3: Search terms

Search Terms

Our search strings have been developed through preliminary database searches to assess search term sensitivity and precision[45]. We also have considered feedback from advisory group members (see Appendix 1) about where to prioritise and how to define cognitive science concepts. The search terms will be based on the interventions outlined in our conceptual map (Appendix 2). For each concept, the string will contain terms related to a) methodology, b) education (outcomes and classroom specific), and c) terms and synonyms related to the specific cognitive science area (including a general cognitive science search). These search terms will be entered into each search database with the minimum of adaptation needed to use the search syntax and functionality and ensure comparability across databases.

Table 3 – General Search Terms (all searches)

Search Term Group	***Search String (Fragment)***	***Search Location[1]***
Group 1 – Methodology	intervention OR trial OR evaluat* OR experiment* OR quasi-experiment* OR pilot OR test*	Title, abstract or key words
Group 2 – Education Outcomes	AND learning OR attainment OR achievement OR "test scores" OR outcomes OR exam* OR impact OR effect OR performance	
Group 3 – Classroom setting	AND classroom OR teach* OR school OR “further education” OR nursery OR "early years" OR kindergarten OR pre-primary OR lesson	
Group 4 – Focus Concept	AND, one of the general or concept-specific search term fragments in Table 4, below.	

[1] Subject to search database functionality

The general search terms above will be combined with one of the search strings related to cognitive science in general and specific cognitive science concepts, below.

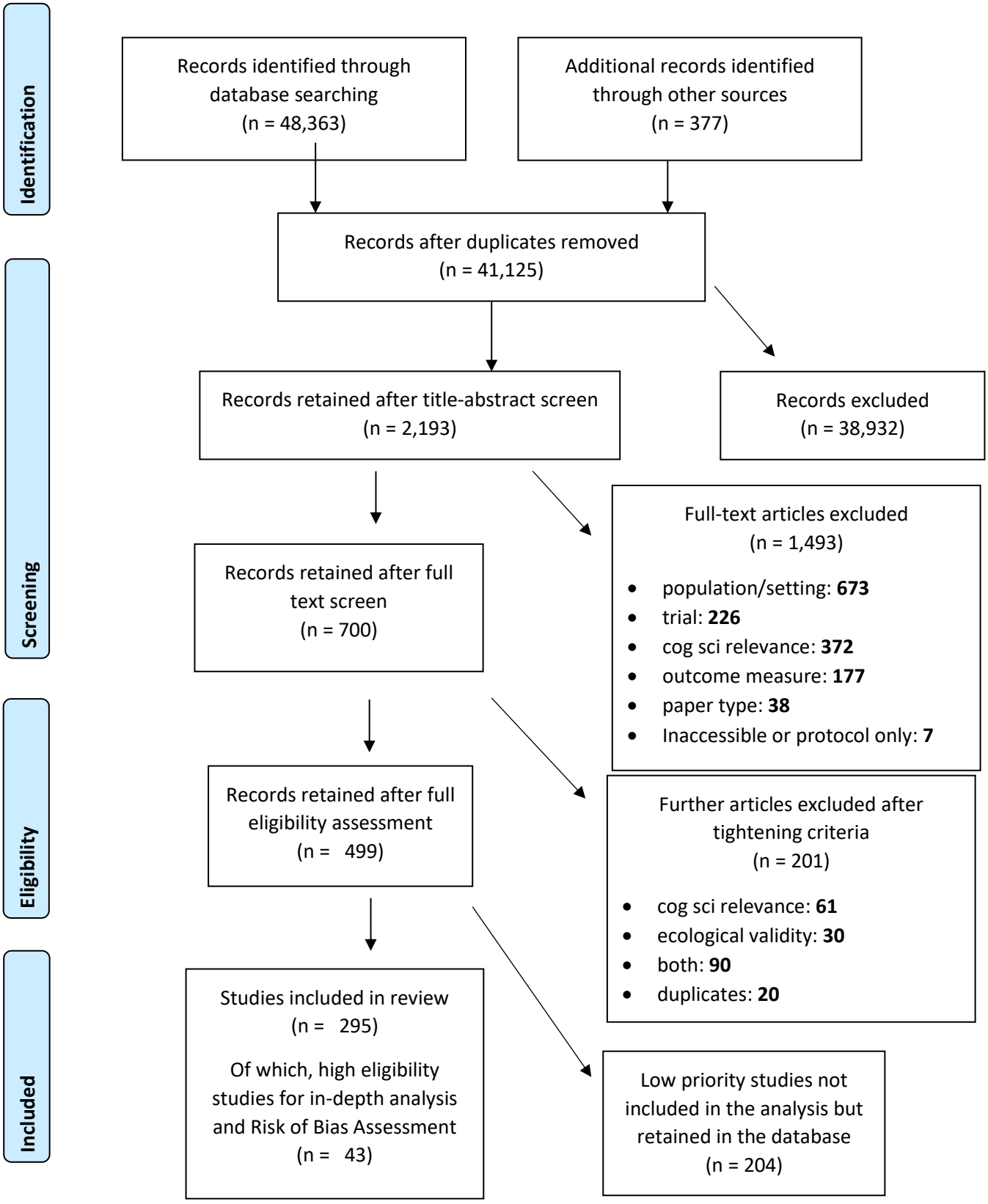
Table 4 – Cognitive Science Concept-specific Search Terms – Core Concepts

Cognitive Science Concept	***Search String (Fragment – to be combined with the general search terms, above)***	***Search Location[1]***
Cognitive Science General	cog* OR brain* OR neuro* OR “learning science”	Title, abstract or key words
Spaced practice	spac* OR distributed	
Interleaving	interleav* OR interweav*	
Retrieval practice	retriev* OR “testing effect”	
Dual coding	dual	
Strategies to manage cognitive load	"working memory" OR "short-term memory" OR (load AND (Cognitive OR intrinsic OR extraneous OR germane))	

[1] Subject to search database functionality

[45] https://handbook-5-1.cochrane.org/chapter_6/6_4_4_sensitivity_versus_precision.htm

Appendix 4: PRISMA flow diagram



Appendix 5: spaced practice

Summary of risk of bias (rob) analysis

		Bias					
Strategy	***Study***	***Randomisation process***	***Deviations from intended intervention***	***Missing outcome data***	***Measurement of the outcome***	***Selection of reported results***	***Overall***
1 - Standard	Feddern et al. (2018)	Low	Low	Low	Low	Low	Low
1 - Standard and 2 - Short	O'Hare et al. (2017)	Low	Low	Low	Low	Low	Low
1 - Standard	Nazari et al. (2019)	Some concerns	Some concerns	Low	Low	Low	Some concerns
2 - Short	Churches et al. (2020)	Some concerns	Some concerns	Low	Some concerns	Low	Some concerns
2 - Short	Kelley et al. (2013)	Some concerns	Some concerns	Low	Low	Some concerns	Some concerns

Database references – 'standard' spacing (across days or lessons)

Short Reference	Focus	Full Reference
Bloom *et al.* (1981)	Effect of spaced practice on retention of second language vocabulary	Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. The Journal of Educational Research, 74(4), 245-248.
Denton *et al.* (2011)	Effects of spacing a group reading intervention on reading outcomes	Denton, C. A., Cirino, P. T., Barth, A. E., Romain, M., Vaughn, S., Wexler, J., ... & Fletcher, J. M. (2011). An experimental study of scheduling and duration of "Tier 2" first-grade reading intervention. Journal of research on educational effectiveness, 4(3), 208-230.
*Feddern *et al.* (2018)	Testing the effectiveness of cognitive science-inspired biology revision software improved biology test scores	Feddern, L., Belham, F. S., & Wilks, S. (2018). Retrieval, interleaving, spacing and visual cues as ways to improve independent learning outcomes at scale. Impact, Journal of the Chartered College of Teaching, 18, 19. Available: https://impact.chartered.college/article/feddern-retrieval-interleaving-spacing-visual-cues-independent-learning/
Foot *et al* (2019)	Effect of spacing on fact-learning and critical thinking	Foot, V. L. (2019). Judging the Credibility of Websites: An Effectiveness Trial of the Spacing Effect in the Elementary Classroom.
French *et al.* (1990)	Effect of spaced practice on volleyball skill	French, K. E., Rink, J. E., & Werner, P. H. (1990). Effects of contextual interference on retention of three volleyball skills. *Perceptual and motor skills*, *71*(1), 179-186.
Goossens *et al.* (2012)	Effect of spaced teaching on vocabulary learning	Goossens, N. A., Camp, G., Verkoeijen, P. P., Tabbers, H. K., & Zwaan, R. A. (2012). Spreading the words: A spacing effect in vocabulary learning. *Journal of Cognitive Psychology*, *24*(8), 965-971.
Goossens *et al.* (2016)^	Effect of retrieval practice and spaced practice on vocabulary learning	Goossens, N. A., Camp, G., Verkoeijen, P. P., Tabbers, H. K., Bouwmeester, S., & Zwaan, R. A. (2016). Distributed practice and retrieval practice in primary school vocabulary learning: A multi-classroom study. *Applied Cognitive Psychology*, *30*(5), 700-712.
Greving & Richter (2019)	Effect of spacing reading on biology text recall and comprehension	Greving, C. E., & Richter, T. (2019). Distributed learning in the classroom: effects of rereading schedules depend on time of test. *Frontiers in psychology*, *9*, 2517.
Kupper-Tetzel *et al.* (2014)	Effect of spaced learning on EFL vocabulary recall	Küpper-Tetzel, C. E., Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science*, *42*(3), 373-388.
Namaziandost *et al.* (2019)	Effect of spaced instruction on EFL vocabulary learning	Namaziandost, E., Nasri, M., Esfahani, F. R., Keshmirshekan, M. H., & Agudo, J. D. D. M. (2019). The impacts of spaced and massed distribution

		instruction on EFL learners' vocabulary learning. *Cogent Education*, *6*(1), 1661131.
*Nazari *et al.* (2019)	Effect of spaced practice on maths (Grade 3 multiplication; Grade 7 probability)	Nazari, K. B., & Ebersbach, M. (2019). Distributed practice in mathematics: Recommendable especially for students on a medium performance level?. *Trends in neuroscience and education*, *17*, 100122.
Nazari & Ebersback (2018)	Effect of homework-based spaced practice on statistics learning	Nazari, K. B., & Ebersbach, M. (2018). Distributed practice: rarely realized in self-regulated mathematical learning. *Frontiers in psychology*, *9*, 2170.
Nazari & Ebersback (2019)	Effect of spaced practice on mathematical performance (probability)	Nazari, K. B., & Ebersbach, M. (2019). Distributing mathematical practice of third and seventh graders: A pplicability of the spacing effect in the classroom. *Applied Cognitive Psychology*, *33*(2), 288-298.
*O'Hare *et al.* (2017)	Evaluation of EEF SMART Spaces programme on GCSE science test performance	O'Hare, L. I. A. M. (2017). Applying the Spacing Effect in the Classroom: The SMART Spaces program. EEF. Available: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/spaced-learning/
Peterson-Brown *et al.* (2019)	Effect of spaced practice on retention of maths vocabulary	Petersen-Brown, S., Lundberg, A. R., Ray, J. E., Dela Paz, I. N., Riss, C. L., & Panahon, C. J. (2019). Applying spaced practice in the schools to teach math vocabulary. *Psychology in the Schools*, *56*(6), 977-991.
Seabrook *et al.* (2005)	Effect of within-session spaced presentation on vocabulary learning and phonics	Seabrook, R., Brown, G. D., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied cognitive psychology*, *19*(1), 107-122.
Sobel *et al.* (2011)	Effect of spaced practice on vocabulary learning	Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*(5), 763-767.
Svihla *et al.* (2018)	Effect of spaced practice on inquiry science learning	Svihla, V., Wester, M. J., & Linn, M. C. (2018). Distributed practice in classroom inquiry science learning. *Learning: Research and Practice*, *4*(2), 180-202.

* High priority study, identified for in-depth analysis; ^ = study included for more than one strategy.

Database references – 'short' spacing (within lessons)

Short Reference	Focus	Full Reference
*Churches *et al.* (2020)^	Teachers designed and led RCTs utilizing cognitive science principles (e.g., spaced practice, retrieval practice, attention), with support from educational neuroscientists	Churches, R., Dommett, E. J., Devonshire, I. M., Hall, R., Higgins, S., & Korin, A. (2020). Translating Laboratory Evidence into Classroom Practice with Teacher-Led Randomized Controlled Trials—A Perspective and Meta-Analysis. *Mind, Brain, and Education*, *14*(3), 292-302.
*Kelley *et al.* (2013)	Effects of spaced learning on biology test scores	Kelley, P., & Whatson, T. (2013). Making long-term memories in minutes: a spaced learning pattern from memory research in education. *Frontiers in human neuroscience*, *7*, 589.
*O'Hare *et al.* (2017)	Evaluation of EEF SMART Spaces programme on GCSE science test performance	O'Hare, L. I. A. M. (2017). Applying the Spacing Effect in the Classroom: The SMART Spaces program. EEF. Available: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/spaced-learning/

* High priority study, identified for in-depth analysis; ^ = study included for more than one strategy.

Database references – wider evidence in this area

Austin, D. A. (1975). Effect of Distributed and Massed Practice upon the Learning of a Velocity Task. Research Quarterly. American Alliance for Health, Physical Education and Recreation, 46(1), 23-30.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. The Journal of Educational Research, 74(4), 245-248.

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Extending cognitive load theory to incorporate working memory resource depletion: evidence from the spacing effect. Educational Psychology Review, 30(2), 483-501.

Chiara, L., Schuster, J. W., Bell, J. K., & Wolery, M. (1995). Small-group massed-trial and individually-distributed-trial instruction with preschoolers. Journal of Early Intervention, 19(3), 203-217.

Codding, R. S., Volpe, R. J., Martin, R. J., & Krebs, G. (2019). Enhancing mathematics fluency: Comparing the spacing of practice sessions with the number of opportunities to respond. School Psychology Review, 48(1), 88-97.

Collins, L., Halter, R. H., Lightbown, P. M., & Spada, N. (1999). Time and the distribution of time in L2 instruction. TESOL quarterly, 33(4), 655-680.

Collins, L., & White, J. (2011). An intensive look at intensity and language learning. Tesol Quarterly, 45(1), 106-133.

Denton, C. A., Cirino, P. T., Barth, A. E., Romain, M., Vaughn, S., Wexler, J., ... & Fletcher, J. M. (2011). An experimental study of scheduling and duration of "Tier 2" first-grade reading intervention. Journal of research on educational effectiveness, 4(3), 208-230.

Fishman, E. J., Keller, L., & Atkinson, R. C. (1968). Massed versus distributed practice in computerized spelling drills. Journal of educational psychology, 59(4), 290.

Feddern, L., Belham, F. S., & Wilks, S. (2018). Retrieval, interleaving, spacing and visual cues as ways to improve independent learning outcomes at scale. Impact, Journal of the Chartered College of Teaching, 18, 19. Available: https://impact.chartered.college/article/feddern-retrieval-interleaving-spacing-visual-cues-independent-learning/

Gettinger, M., Bryant, N. D., & Fayne, H. R. (1982). Designing spelling instruction for learning-disabled children: An emphasis on unit size, distributed practice, and training for transfer. The Journal of Special Education, 16(4), 439-448.

Gluckman, M., Vlach, H. A., & Sandhofer, C. M. (2014). Spacing simultaneously promotes multiple forms of learning in children's science curriculum. Applied Cognitive Psychology, 28(2), 266-273.

Grassi, J. R. (1971). Effects of massed and spaced practice on learning in brain-damaged, behavior-disordered, and normal children. Journal of Learning Disabilities, 4(5), 237-242.

Griffin, C., & Joseph, L. M. (2015). Supplemental Flashcard Drill Methods for Efficiently Helping At-Risk Kindergartners Make Letter-Sound Correspondences: Does Presentation Arrangement of Words Matter?. Reading Psychology, 36(5), 421-444.

Kasprowicz, R. E., Marsden, E., & Sephton, N. (2019). Investigating distribution of practice effects for the learning of foreign language verb morphology in the young learner classroom. The Modern Language Journal, 103(3), 580-606.

Krug, D., Davis, T. B., & Glover, J. A. (1990). Massed versus distributed repeated reading: A case of forgetting helping recall?. Journal of Educational Psychology, 82(2), 366.

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. Psychological science, 25(3), 639-647.

Lotfolahi, A. R., & Salehi, H. (2017). Spacing effects in vocabulary learning: Young EFL learners in focus. Cogent Education, 4(1), 1287391.

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based programme to enhance study efficacy in a high and low risk setting. European Journal of Cognitive Psychology, 19(4-5), 743-768.

Nowbakht, M., Moinzadeh, A., & Dabaghi, A. (2015). Effects of Massed vs. Distributed Implicit FonF on Receptive Acquisition of L2 Vocabulary Items. Journal of Asia TEFL, 12(3).

Schutte, G. M., Duhon, G. J., Solomon, B. G., Poncy, B. C., Moore, K., & Story, B. (2015). A comparative analysis of massed vs. distributed practice on basic math fact fluency growth rates. Journal of School Psychology, 53(2), 149-159.

Snoder, P. (2017). Improving English Learners' Productive Collocation Knowledge: The Effects of Involvement Load, Spacing, and Intentionality. TESL Canada Journal, 34(3), 140-164.

Stambaugh, L. A. (2011). When repetition isn't the best practice strategy: Effects of blocked and random practice schedules. Journal of Research in Music Education, 58(4), 368-383.

Stoltzfus, M., & Sukseemuang, P. (2018). Distribution of Instructional Time in Secondary, Non-Intensive Thai EFL Classes: Effects on Grammar Acquisition. Electronic Journal of Foreign Language Teaching, 15(2).

Swehla, S. E., Burns, M. K., Zaslofsky, A. F., Hall, M. S., Varma, S., & Volpe, R. J. (2016). Examining the use of spacing effect to increase the efficiency of incremental rehearsal. Psychology in the Schools, 53(4), 404-415.

Veal, W. R. (1999). What could define block scheduling as a fad?. American Secondary Education, 27(4), 3-12.

Vlach, H. A., & Sandhofer, C. M. (2012). Distributing learning over time: The spacing effect in children's acquisition and generalization of science concepts. Child development, 83(4), 1137-1144.

Zigterman, J. R., Simone, P. M., & Bell, M. C. (2015). Within-session spacing improves delayed recall in children. Memory, 23(4), 625-632.

Appendix 6: interleaving

Summary of risk of bias (rob) analysis

	Bias					
Study	***Randomisation process***	***Deviations from intended intervention***	***Missing outcome data***	***Measurement of the outcome***	***Selection of reported results***	***Overall***
Booth *et al*. (2015)	Some concerns	Low	Some concerns	Low	Some concerns	Some concerns
Nemeth *et al*. (2019)	Some concerns	Low	Some concerns	Low	Some concerns	Some concerns
Rau *et al*. (2013)	Some concerns	Low	Some concerns	Low	Some concerns	Some concerns
Rohrer *et al*. (2014)	Some concerns	Low	Low	Low	Some concerns	Some concerns
Rohrer *et al*. (2015)	Some concerns	Low	Low	Low	Some concerns	Some concerns
Rohrer *et al*. (2019)	Low	Low	Low	Low	Low	Low

Database references – interleaving

Short Reference	Focus	Full Reference
*Booth *et al*. (2015a)	Effect of AlgebraByExample assignments on algebra test scores	Booth, J. L., Cooper, L. A., Donovan, M. S., Huyghe, A., Koedinger, K. R., & Paré-Blagoev, E. J. (2015). Design-based research within the constraints of practice: AlgebraByExample. Journal of Education for Students Placed at Risk (JESPAR), 20(1-2), 79-100.
French *et al*. (1990)	Effect of spaced practice on volleyball skill	French, K. E., Rink, J. E., & Werner, P. H. (1990). Effects of contextual interference on retention of three volleyball skills. Perceptual and motor skills, 71(1), 179-186.
*Nemeth *et al*. (2019)	Flexible use of algorithmic and number strategies in elementary school maths (subtraction)	Nemeth, L., Werker, K., Arend, J., Vogel, S., & Lipowsky, F. (2019). Interleaved learning in elementary school mathematics: Effects on the flexible and adaptive use of subtraction strategies. Frontiers in psychology, 10, 86.
Patel *et al*. (2016)	Interleaving versus blocking fraction addition and multiplication practice	Patel, R., Liu, R., & Koedinger, K. R. (2016). When to Block versus Interleave Practice? Evidence Against Teaching Fraction Addition before Fraction Multiplication. In CogSci.
*Rau *et al*. (2013)	The effect of interleaving multiple representations versus tasks types for fractions learning.	Rau, M. A., Aleven, V., & Rummel, N. (2013). Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave?. Learning and Instruction, 23, 98-114.
Rau *et al*. (2014)	Interleaved versus blocked sequences of multiple representation of fractions.	Rau, M. A., Aleven, V., Rummel, N., & Pardos, Z. (2014). How should intelligent tutoring systems sequence multiple graphical representations of fractions? A multi-methods study. International Journal of Artificial Intelligence in Education, 24(2), 125-161.
*Rohrer *et al*. (2014)	Interleaving in mathematics task types requiring strategy selection	Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. Psychonomic bulletin & review, 21(5), 1323-1330.
*Rohrer *et al*. (2015)	Interleaving in mathematics task types requiring strategy selection	Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. Journal of Educational Psychology, 107(3), 900.
*Rohrer *et al*. (2019)	Interleaving in mathematics task types requiring strategy selection	Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C. N. (2020). A randomized controlled trial of interleaved mathematics practice. Journal of Educational Psychology, 112(1), 40.

Todaro *et al*. (2017)	Interleaving geometry problems and contexts	Todaro, R., & Morris, B. J. (2017). Interleaving area problems in the 4th grade classroom: What is the role of context and practice?. In CogSci.
Todaro *et al*. (2019)	Contextual, concrete, or abstract example manipulations in interleaved vs. blocked sequences in maths	Todaro, R. D. (2019). Investigating the Role of Example Type in Interleaved Practice (Doctoral dissertation, Kent State University).
Wagner *et al*. (2019)	Effect of interleaved practice vs. repetitive practice and incremental rehearsal when learning single digit addition and multiplication facts	Wagner, K. (2019). Examination of Three Practice Schedules for Single Digit Math.

* High priority study, identified for in-depth analysis

Database references – wider evidence in this area

Mandler, J. M., & DeForest, M. (1979). Is there more than one way to recall a story?. Child Development, 886-889.

Rittle-Johnson, B., & Koedinger, K. (2009). Iterating between lessons on concepts and procedures can improve mathematics knowledge. British Journal of Educational Psychology, 79(3), 483-500.

Rittle-Johnson, B., & Koedinger, K. R. (2002). Comparing Instructional Strategies for Integrating Conceptual and Procedural Knowledge.

Ziegler, E., Edelsbrunner, P. A., & Stern, E. (2018). The relative merits of explicit and implicit learning of contrasted algebra principles. Educational Psychology Review, 30(2), 531-558.

Appendix 7: retrieval practice

Summary of risk of bias (rob) analysis

	Bias					
Study	***Randomisation process***	***Deviations from intended intervention***	***Missing outcome data***	***Measurement of the outcome***	***Selection of reported results***	***Overall***
Agarwal (2019)	Some concerns	Low	Low	Low	Some concerns	Some concerns
Churches *et al.* (2020)	Some concerns	Some concerns	Low	Some concerns	Low	Some concerns
Damhuis *et al.* (2016)[1]	High (not randomised)	Low	Low	Low	Some concerns	High
Roediger *et al.* (2011)	Low	Low	Low	Low	Some concerns	Some concerns

[1] n.b. this study tested adaptive vs. regular retrieval practice and is included in the wider evidence rather than the main strategy evidence review.

Database references – retrieval practice (testing effect)

Short Reference	Focus	Full Reference
*Agarwal (2019): Expt.3 only	Effect of retrieval practice and question type on higher-order learning in history	Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning?. Journal of Educational Psychology, 111(2), 189.
Barenberg and Dutke (2019)	Effect of retrieval practice on comprehension accuracy and confidence in judgements	Barenberg, J., & Dutke, S. (2019). Testing and metacognition: retrieval practise effects on metacognitive monitoring in learning from text. Memory, 27(3), 269-279.
Carpenter *et al.* (2009)	Effect of review/testing on recall of US history facts	Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 23(6), 760-771.
*Churches *et al.* (2020)^	Teachers designed and led RCTs utilizing cognitive science principles (e.g., spaced practice, retrieval practice, attention), with support from educational neuroscientists	Churches, R., Dommett, E. J., Devonshire, I. M., Hall, R., Higgins, S., & Korin, A. (2020). Translating Laboratory Evidence into Classroom Practice with Teacher-Led Randomized Controlled Trials—A Perspective and Meta-Analysis. Mind, Brain, and Education, 14(3), 292-302.
Damhuis *et al.* (2015)	Effects of repeated storybook reading versus testing on vocabulary learning	Damhuis, C. M., Segers, E., & Verhoeven, L. (2015). Stimulating breadth and depth of vocabulary via repeated storybook readings or tests. School Effectiveness and School Improvement, 26(3), 382-396.
Dirkx *et al.* (2014)	Effect of testing on learning from principles and procedures from texts	Dirkx, K. J., Kester, L., & Kirschner, P. A. (2014). The testing effect for learning principles and procedures from texts. The Journal of Educational Research, 107(5), 357-364.
Goossens *et al.* (2014a)	Effect of retrieval practice and learning context on vocabulary learning	Goossens, N. A., Camp, G., Verkoeijen, P. P., & Tabbers, H. K. (2014). The effect of retrieval practice in primary school vocabulary learning. Applied Cognitive Psychology, 28(1), 135-142.
Goossens *et al.* (2014b)	Effect of retrieval practice on vocabulary learning	Goossens, N. A., Camp, G., Verkoeijen, P. P., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. Journal of Applied Research in Memory and Cognition, 3(3), 177-182.

Goossens *et al*. (2016)^	Effect of retrieval practice and spaced practice on vocabulary learning	Goossens, N. A., Camp, G., Verkoeijen, P. P., Tabbers, H. K., Bouwmeester, S., & Zwaan, R. A. (2016). Distributed practice and retrieval practice in primary school vocabulary learning: A multi-classroom study. Applied Cognitive Psychology, 30(5), 700-712.
Hanham *et al*. (2017)	Effect of testing and element interactivity (complexity) on learning to write types of text	Hanham, J., Leahy, W., & Sweller, J. (2017). Cognitive load theory, element interactivity, and the testing and reverse testing effects. Applied Cognitive Psychology, 31(3), 265-280.
Jaeger *et al*. (2015)	Effect of retrieval practice on recall of information from texts	Jaeger, A., Eisenkraemer, R. E., & Stein, L. M. (2015). Test-enhanced learning in third-grade children. Educational Psychology, 35(4), 513-521.
Jagerskog *et al*. (2019)^	Effect of retrieval practice versus multimedia learning on psychology recall	Jägerskog, A. S., Jönsson, F. U., Selander, S., & Jonsson, B. (2019). Multimedia learning trumps retrieval practice in psychology teaching. Scandinavian journal of psychology, 60(3), 222-230.
Karpicke *et al*. (2014)	Effect of retrieval practice on recall of science texts	Karpicke, J. D., Blunt, J. R., Smith, M. A., & Karpicke, S. S. (2014). Retrieval-based learning: The need for guided retrieval in elementary school children. Journal of Applied Research in Memory and Cognition, 3(3), 198-206.
Karpicke *et al*. (2016)	Effect of retrieval practice on word recall	Karpicke, J. D., Blunt, J. R., & Smith, M. A. (2016). Retrieval-based learning: Positive effects of retrieval practice in elementary school children. Frontiers in Psychology, 7, 350.
Lipowski *et al*. (2014)	Effect of testing on word recall	Lipowski, S. L., Pyc, M. A., Dunlosky, J., & Rawson, K. A. (2014). Establishing and explaining the testing effect in free recall for young children. Developmental Psychology, 50(4), 994.
McDaniel *et al*. (2011)	Effect of quiz frequency and placement on science test scores	McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger III, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. Journal of Educational Psychology, 103(2), 399.
McDermott *et al*. (2014)	Effect of quiz type on history and science test scores	McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger III, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. Journal of Experimental Psychology: Applied, 20(1), 3.
Nungester and Duchastel (1982)	Effect of testing on retention of history knowledge	Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. Journal of Educational Psychology, 74(1), 18.
Ritchie *et al*. (2013)	Effect of retrieval practice (with or without mind-mapping) on geographical fact learning	Ritchie, S. J., Della Sala, S., & McIntosh, R. D. (2013). Retrieval practice, with or without mind mapping, boosts fact learning in primary school children. PloS one, 8(11), e78976.
*Roediger *et al*. (2011)	Effect of quizzing on social study test scores	Roediger III, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. Journal of Experimental Psychology: Applied, 17(4), 382.
Urhahne *et al*. (2013)	Effect of retrieval practice task type on science knowledge	Urhahne, D., Nick, S., Poepping, A. C., & Schulz, S. J. (2013). The effects of study tasks in a computer-based chemistry learning environment. Journal of Science Education and Technology, 22(6), 993-1003.

* High priority study, identified for in-depth analysis; ^ = study included for more than one strategy.

Database references – wider evidence in this area

Aslan, A., & Bäuml, K. H. T. (2016). Testing enhances subsequent learning in older but not in younger elementary school children. Developmental Science, 19(6), 992-998.

Bebko, J. M., Rhee, T., Ncube, B. L., & Dahary, H. (2017). Effectiveness and retention of teaching memory strategy use to children with autism spectrum disorder. Canadian Journal of School Psychology, 32(3-4), 244-264.

Bertilsson, F., Wiklund-Hörnqvist, C., Stenlund, T., & Jonsson, B. (2017). The testing effect and its relation to working memory capacity and personality characteristics. Journal of Cognitive Education and Psychology, 16(3), 241-259.

Bouwmeester, S., & Verkoeijen, P. P. (2011). The effect of instruction method and relearning on Dutch spelling performance of third-through fifth-graders. European journal of psychology of education, 26(1), 61-74.

Bouwmeester, S., & Verkoeijen, P. P. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. Journal of Memory and Language, 65(1), 32-41.

Ramirez Butavand, D., Hirsch, I., Tomaiuolo, M., Moncada, D., Viola, H., & Ballarini, F. (2020). Novelty improves the formation and persistence of memory in a naturalistic school scenario. Frontiers in psychology, 11, 48.

Carneiro, P., Lapa, A., & Finn, B. (2018). The effect of unsuccessful retrieval on children's subsequent learning. Journal of experimental child psychology, 166, 400-420.
Cassaday, H. J., Bloomfield, R. E., & Hayward, N. (2002). Relaxed conditions can provide memory cues in both undergraduates and primary school children. British Journal of educational psychology, 72(4), 531-547.
Chang, C. Y., Yeh, T. K., & Barufaldi, J. P. (2010). The positive and negative effects of science concept tests on student conceptual understanding. International Journal of Science Education, 32(2), 265-282.
Chen, C. H., & Huang, K. (2014). The effects of response modes and cues on language learning, cognitive load and self-efficacy beliefs in web-based learning. Journal of Educational Multimedia and Hypermedia, 23(2), 117-134.
Cooper, M. H., & Newman, S. E. (1987). Cue-target compatibility in children's cued recall. The American journal of psychology, 167-178.
Damhuis, C. M., Segers, E., Scheltinga, F., & Verhoeven, L. (2016). Effects of individualized word retrieval in kindergarten vocabulary intervention. School Effectiveness and School Improvement, 27(3), 441-454.
Davis, Z. T. (1987). Effects of time-of-day of instruction on beginning reading achievement. The Journal of Educational Research, 80(3), 138-140.
Duchastel, P. C. (1979). Retention of prose materials: The effect of testing. The Journal of Educational Research, 72(5), 299-300.
Duchastel, P., & Nungester, R. (1981). Long-term retention of prose following testing. Psychological Reports.
Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. Contemporary Educational Psychology, 6(3), 217-226.
Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. The Journal of Educational Research, 75(5), 309-313.
Folkard, S., Monk, T. H., Bradbury, R., & Rosenthall, J. (1977). Time of day effects in school children's immediate and delayed recall of meaningful material. British journal of Psychology, 68(1), 45-50.
Griffin, C., & Joseph, L. M. (2015). Supplemental Flashcard Drill Methods for Efficiently Helping At-Risk Kindergartners Make Letter-Sound Correspondences: Does Presentation Arrangement of Words Matter?. Reading Psychology, 36(5), 421-444.
Guza, D. S., & McLaughlin, T. F. (1987). A comparison of daily and weekly testing on student spelling performance. The Journal of Educational Research, 80(6), 373-376.
Howe, M. L. (2002). The role of intentional forgetting in reducing children's retroactive interference. Developmental Psychology, 38(1), 3.
Hwang, Y., & Levin, J. R. (2002). Examination of middle-school students' independent use of a complex mnemonic system. The Journal of experimental education, 71(1), 25-38.
Imuta, K., Scarf, D., Carson, S., & Hayne, H. (2018). Children's learning and memory of an interactive science lesson: Does the context matter?. Developmental psychology, 54(6), 1029.
Jones, A. C., Wardlow, L., Pan, S. C., Zepeda, C., Heyman, G. D., Dunlosky, J., & Rickard, T. C. (2016). Beyond the rainbow: Retrieval practice leads to better spelling than does rainbow writing. Educational Psychology Review, 28(2), 385-400.
Kliegl, O., Abel, M., & Bäuml, K. H. T. (2018). A (preliminary) recipe for obtaining a testing effect in preschool children: Two critical ingredients. Frontiers in psychology, 9, 1446.
Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. Educational Psychology Review, 27(2), 291-304.
Leggett, J. M., Burt, J. S., & Carroll, A. (2019). Retrieval practice can improve classroom review despite low practice test performance. Applied Cognitive Psychology, 33(5), 759-770.
Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science?. Journal of Applied Research in Memory and Cognition, 3(3), 171-176.
Marsh, E. J., Fazio, L. K., & Goswick, A. E. (2012). Memorial consequences of testing school-aged children. Memory, 20(8), 899-906.
Mateo, A., Ros, L., Ricarte, J. J., Fernandez, D., & Latorre, J. M. (2020). Effects of visual and verbal cues in facilitating the remembering of an autobiographical event in preschoolers. Early Child Development and Care, 190(7), 1093-1108.
McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. Applied Cognitive Psychology, 27(3), 360-372.
Mok, W. S. Y., & Chan, W. W. L. (2016). How do tests and summary writing tasks enhance long-term retention of students with different levels of test anxiety?. Instructional Science, 44(6), 567-581.
Pachman, M., Sweller, J., & Kalyuga, S. (2013). Levels of knowledge and deliberate practice. Journal of Experimental Psychology: Applied, 19(2), 108.
Pals, F. F., Tolboom, J. L., Suhre, C. J., & van Geert, P. L. (2018). Memorisation methods in science education: tactics to improve the teaching and learning practice. International Journal of Science Education, 40(2), 227-241.
Roelle, J., Roelle, D., & Berthold, K. (2019). Test-based learning: inconsistent effects between higher-and lower-level test questions. The Journal of Experimental Education, 87(2), 299-313.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36(1), 233.

Rowley, T., & McCrudden, M. T. (2020). Retrieval practice and retention of course content in a middle school science classroom. Applied Cognitive Psychology, 34(6), 1510-1515.

Scruggs, T. E., Mastropieri, M. A., Brigham, F. J., & Sullivan, G. S. (1992). Effects of mnemonic reconstructions on the spatial learning of adolescents with learning disabilities. Learning Disability Quarterly, 15(3), 154-162.

Stenlund, T., Jönsson, F. U., & Jonsson, B. (2017). Group discussions and test-enhanced learning: individual learning outcomes and personality characteristics. Educational Psychology, 37(2), 145-156.

Tornare, E., Cuisinier, F., Czajkowski, N. O., & Pons, F. (2017). Impact of induced joy on literacy in children: does the nature of the task make a difference?. Cognition and emotion, 31(3), 500-510.

Beyza, U. Ç. A. R., & ÇEVİK, Y. D. Examination of Different Learning Conditions in the Testing Effect: Example of Safe Internet Use. Bartın Üniversitesi Eğitim Fakültesi Dergisi, 7(1), 29-66.

Van den Broek, G. S., Segers, E., Van Rijn, H., Takashima, A., & Verhoeven, L. (2019). Effects of elaborate feedback during practice tests: Costs and benefits of retrieval prompts. Journal of Experimental Psychology: Applied, 25(4), 588.

Wetzels, S. A., Kester, L., van Merriënboer, J. J., & Broers, N. J. (2011). The influence of prior knowledge on the retrieval-directed function of note taking in prior knowledge activation. British Journal of Educational Psychology, 81(2), 274-291.

Retrieval practice studies in churches *et al*. (2020)

Baker, S., & Hindley, C. (2018a, March). Combining retrieval practices with look, cover, write, Check improves pupils' progress in spelling. Poster session presented at Third Space: Science of Learning Conference, Chartered College of Teaching, University of York, York.

Baker, S., & Hindley, C. (2018b, June). Retrieval practices (multiple choice testing) and recall of multiplication tables – A year 4 trial and parallel year 5 replication. Poster session presented at European Conference on Education, IAFOR, Brighton.

Dunford, E., & Rhoades, T. (2018, September). Using an app that provides 10 minutes of retrieval practice every day improves the speed of times table recall in a rural primary school in England – A preliminary study with parallel year 2 and year 3 replications. Poster session presented at BERA Conference, Northumbria University, Newcastle.

Elliott, R., & Wyatt, R. (2018, March). Using retrieval practice improves pupil progress in times table tests with year 4 pupils – A randomised controlled trial. Poster session presented at EARLI SIG22, Neuroscience and Education Conference, Wellcome Trust, London, England.

Greenfield, J., Noden, J., & Siddle, J. (2018, September). The effect of interleaving and retrieval practice on times table retention a randomised controlled trial. Poster session presented at BERA Conference, Northumbria University, Newcastle.

Maberly, L. (2018, March). A preliminary study into the effectiveness of using revision flashcards to improve performance in GCSE end-of-topic assessment. Poster session presented at Third Space: Science of Learning Conference, Chartered College of Teaching, University of York, York.

Makarova, D. (2018, October). A small-scale randomised controlled trial into the effect of practice testing on attainment in KS4 Science. Poster session presented at International Mind, Brain, and Education Society Conference, University of Southern California, Los Angeles.

Morris, C. (2018, March). Retrieval practice, the use of novelty and normal classroom practice – A small-scale randomised controlled trial using a 2 × 2 factorial design, in an English rural primary school. Poster session presented at Academy of Principals, 9th Global Educational Leadership Conference, Singapore.

Pemberton, R. (2018, September). Preliminary evidence for the effectiveness of software based on retrieval practice, compared with peer-based learning and current ordinary school practice. A pilot within-participant design in a small rural primary school. Poster session presented at BERA Conference, Northumbria University, Newcastle.

Quinn, L., & Lamb, J. (2018, March). The effect of retrieval practice on retention and recall of vocabulary with year 8 English students. Poster session presented at Third Space: Science of Learning Conference, Chartered College of Teaching, University of York, York.

Siddle, J. (2018, October). The effect of two types of retrieval practice on vocabulary recall with different aged primary school pupils – A randomised controlled trial incorporating four parallel replications (EYFS, year 2, year 3 and year 5). Poster session presented at International Mind, Brain, and Education Society Conference, University of Southern California, Los Angeles

Appendix 8: Working with Schemas

Summary of risk of bias (rob) analysis

		Bias					
Strategy	**Study**	***Randomisation process***	***Deviations from intended intervention***	***Missing outcome data***	***Measurement of the outcome***	***Selection of reported results***	***Overall***
Concept-mapping and organisation	Merchie *et al.* (2016)	Low	Low	Low	Low	Some concerns	Some concerns
Concept-mapping and organisation	Milenkovic *et al.* (2014)	Some concerns	Low	Some concerns	Low	Some concerns	Some concerns
Concept-mapping and organisation	Ponce *et al.* (2013)	Low	High	Low	Low	Some concerns	High
Comparison and cognitive conflict	Star *et al.* (2015)	Low	High	High	Low	Some concerns	High

Database references – concept-mapping and organisation

Short Reference	Focus	Full Reference
Chang *et al.* (2002)	The Effect of Concept Mapping to Enhance Text comprehension And Summarization	Chang, K. E., Sung, Y. T., & Chen, I. D. (2002). The effect of concept mapping to enhance text comprehension and summarization. The Journal of Experimental Education, 71(1), 5-23.
Fuchs *et al.* (2004)	Effects of schema-based transfer instruction on real-life mathematical problem-solving	Fuchs, L. S., Fuchs, D., Finelli, R., Courey, S. J., & Hamlett, C. L. (2004). Expanding schema-based transfer instruction to help third graders solve real-life mathematical problems. American Educational Research Journal, 41(2), 419-445.
Guastello *et al.* (2000)	Concept Mapping Effects on Science on content Comprehension of Low-Achieving Inner-City Seventh Graders	Guastello, E. F., Beasley, T. M., & Sinatra, R. C. (2000). Concept mapping effects on science content comprehension of low-achieving inner-city seventh graders. Remedial and special education, 21(6), 356-364.
Hilbert and Renkl. (2009) (Expt.2)	Computer-based concept-mapping tool: Self-explaining examples on marketing text	Hilbert, T. S., & Renkl, A. (2009). Learning how to use a computer-based concept-mapping tool: Self-explaining examples helps. Computers in Human Behavior, 25(2), 267-274.
Jitendra *et al.* (2009)	Effect of schema-based instruction on mathematical problem-solving	Im, S. H., & Jitendra, A. K. (2020). Analysis of proportional reasoning and misconceptions among students with mathematical learning disabilities. The Journal of Mathematical Behavior, 57, 100753. Jaeger, A. J., & Wiley, J. (2015). Reading an analogy can cause the illusion of comprehension. Discourse Processes, 52(5-6), 376-405.
Karpicke *et al.* (2014)	Effect of retrieval practice on recall of science texts	Karpicke, J. D., Blunt, J. R., Smith, M. A., & Karpicke, S. S. (2014). Retrieval-based learning: The need for guided retrieval in elementary school children. Journal of Applied Research in Memory and Cognition, 3(3), 198-206.
Merchie and Van Keer. (2016)*	effectiveness of two instructional approaches of mind mapping used as a metalearning strategy	Merchie, E., & Van Keer, H. (2016). Mind mapping as a meta-learning strategy: Stimulating pre-adolescents' text-learning strategies and performance?. Contemporary Educational Psychology, 46, 128-147.
Milenkovic *et al.* (2014)*	Instructional Strategy Based on the Interaction of Multiple Levels of Knowledge Representation	Milenković, D. D., Segedinac, M. D., & Hrin, T. N. (2014). Increasing high school students' chemistry performance and reducing cognitive load through an instructional strategy based on the interaction of multiple levels of knowledge representation. Journal of Chemical Education, 91(9), 1409-1416.

Okebukola *et al.* (1992)	Individual and collaborative concept-mapping	Okebukola, P. A. (1992). Concept mapping with a cooperative learning flavor. The American Biology Teacher, 218-221.
Ponce *et al.* (2013)*	scaffolded practice in translating passages into graphic organizers	Ponce, H. R., Mayer, R. E., & Lopez, M. J. (2013). A computer-based spatial learning strategy approach that improves reading comprehension and writing. Educational Technology Research and Development, 61(5), 819-840.
Ponce *et al.* (2018)	Computer-supported learning strategies (inc. graphic organisation)	Ponce, H. R., Mayer, R. E., Loyola, M. S., López, M. J., & Méndez, E. E. (2018). When two computer-supported learning strategies are better than one: An eye-tracking study. Computers & Education, 125, 376-388.
Ritchie *et al.* (2013)	Effect of retrieval practice (with or without mind-mapping) on geographical fact learning	Ritchie, S. J., Della Sala, S., & McIntosh, R. D. (2013). Retrieval practice, with or without mind mapping, boosts fact learning in primary school children. PloS one, 8(11), e78976.
Romance *et al.* (2017)	multi-year effects of the Science IDEAS model on science and reading comprehension achievement	Romance, N., & Vitale, M. (2017). Implications of a cognitive science model integrating literacy in science on achievement in science and reading: Direct effects in grades 3–5 with transfer to grades 6–7. International Journal of Science and Mathematics Education, 15(6), 979-995.
Wijekumar *et al.* (2012)	Effect of intelligent tutoring on nonfiction reading comprehension	Wijekumar, K. K., Meyer, B. J., & Lei, P. (2012). Large-scale randomized controlled trial with 4th graders using intelligent tutoring of the structure strategy to improve nonfiction reading comprehension. Educational Technology Research and Development, 60(6), 987-1013.
Wijekumar *et al.* (2017)	Effect of intelligent tutoring on recall of expository texts	Wijekumar, K., Meyer, B. J., Lei, P., Cheng, W., Ji, X., & Joshi, R. M. (2017). Evidence of an intelligent tutoring system as a mindtool to promote strategic memory of expository texts and comprehension with children in grades 4 and 5. Journal of Educational Computing Research, 55(7), 1022-1048.

* High priority study, identified for in-depth analysis

Database references – comparison and cognitive conflict

Short Reference	Focus	Full Reference
Adey and Shayer (1993)	Lessons based on concrete activities, cognitive conflict, metacognition, schema development (bridging[46] of thinking strategies) in science.	Adey, P., & Shayer, M. (1993). An exploration of long-term far-transfer effects following an extended intervention program in the high school science curriculum. Cognition and instruction, 11(1), 1-29.
Chiu and Churchill (2016)	Conceptual variation in algebra teaching, seeing and experiencing different algebraic forms and solving methods simultaneously	Chiu, T. K., & Churchill, D. (2016). Design of learning objects for concept learning: Effects of multimedia learning principles and an instructional approach. Interactive Learning Environments, 24(6), 1355-1370.
Day *et al.* (2015)	use of concrete, familiar examples in science	Day, S. B., Motz, B. A., & Goldstone, R. L. (2015). The cognitive costs of context: The effects of concreteness and immersiveness in instructional examples. Frontiers in Psychology, 6, 1876.
Madu *et al.* (2015)	cognitive-conflict-based physics instruction over the traditionally designed physics instruction on students' conceptual change in heat and temperature	Madu, B. C., & Orji, E. (2015). Effects of cognitive conflict instructional strategy on students' conceptual change in temperature and heat. Sage Open, 5(3), 2158244015594662.
Poehnl *et al.* (2013)	computer and textbook instruction with involving alternative conceptions (ACs)	Poehnl, S., & Bogner, F. X. (2013). Cognitive load and alternative conceptions in learning genetics: Effects from provoking confusion. The Journal of Educational Research, 106(3), 183-196.
Rittle-Johnson *et al.* (2009)	solving equations, comparing different problem types solved with the same solution method, or	Rittle-Johnson, B., & Star, J. R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and

[46] i.e., strategies to generalise reasoning to promote transfer.

	different solution methods to the same problem	procedural flexibility for equation solving. Journal of Educational Psychology, 101(3), 529.
Star *et al*. (2015)*	effect of an Algebra supplemental comparison curriculum on students' mathematical knowledge	Star, J. R., Pollack, C., Durkin, K., Rittle-Johnson, B., Lynch, K., Newton, K., & Gogolen, C. (2015). Learning from comparison in algebra. Contemporary Educational Psychology, 40, 41-54.
Stara *et al*. (2009)	comparison in a classroom context for children learning about computational estimation	Stara, J. R., & Rittle-Johnson, B. (2009). It pays to compare: An experimental study on computational estimation. Journal of Experimental Child Psychology, 102(4), 408-426.
Ziegler and Stern (2014)	learning elementary algebraic transformations through contrasted comparisons	Ziegler, E., & Stern, E. (2014). Delayed benefits of learning elementary algebraic transformations through contrasted comparisons. Learning and Instruction, 33, 131-146.
Ziegler and Stern (2016)	Algebra learning using contrasted comparisons	Ziegler, E., & Stern, E. (2016). Consistent advantages of contrasted comparisons: Algebra learning under direct instruction. Learning and Instruction, 41, 41-51.

* High priority study, identified for in-depth analysis

Database references – wider evidence in this area

Adey, P., Robertson, A., & Venville, G. (2002). Effects of a cognitive acceleration programme on Year I pupils. British Journal of Educational Psychology, 72(1), 1-25.

Barkl, S., Porter, A., & Ginns, P. (2012). Cognitive training for children: Effects on inductive reasoning, deductive reasoning, and mathematics achievement in an Australian school setting. Psychology in the Schools, 49(9), 828-842.

Berthold, K., & Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. Journal of educational psychology, 101(1), 70.

Bramwell-Lalor, S., & Rainford, M. (2014). The effects of using concept mapping for improving advanced level biology students' lower-and higher-order cognitive skills. International Journal of Science Education, 36(5), 839-864.

Butcher, K. R., & Aleven, V. (2013). Using student interactions to foster rule–diagram mapping during problem solving in an intelligent tutoring system. Journal of Educational Psychology, 105(4), 988.

Calin-Jageman, R. J., & Horn Ratner, H. (2005). The role of encoding in the self-explanation effect. Cognition and Instruction, 23(4), 523-543.

Candry, S., Decloedt, J., & Eyckmans, J. (2020). Comparing the merits of word writing and retrieval practice for L2 vocabulary learning. System, 89, 102206.

Casteleyn, J., & Mottart, A. (2012). Presenting material via graphic organizers in science classes in secondary education. Procedia-Social and Behavioral Sciences, 69, 458-466.

Chang, C. C., Liu, G. Y., Chen, K. J., Huang, C. H., Lai, Y. M., & Yeh, T. K. (2017). The effects of a collaborative computer-based concept mapping strategy on geographic science performance in junior high school students. Eurasia Journal of Mathematics, Science and Technology Education, 13(8), 5049-5060.

De Corte, E., Verschaffel, L., & Van De Ven, A. (2001). Improving text comprehension strategies in upper primary school children: A design experiment. British Journal of Educational Psychology, 71(4), 531-559.

Deák, G. O., & Toney, A. J. (2013). Young children's fast mapping and generalization of words, facts, and pictograms. Journal of Experimental Child Psychology, 115(2), 273-296.

Denessen, E., Veenman, S., Dobbelsteen, J., & Van Schilt, J. (2008). Dyad composition effects on cognitive elaboration and student achievement. The Journal of Experimental Education, 76(4), 363-386.

Dincer, S. (2011). Exploring the Impacts of Analogies on Computer Hardware. Turkish Online Journal of Educational Technology-TOJET, 10(2), 113-121.

Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. Quarterly Journal of Experimental Psychology, 60(7), 991-1004.

Fuchs, L. S., Malone, A. S., Schumacher, R. F., Namkung, J., Hamlett, C. L., Jordan, N. C., ... & Changas, P. (2016). Supported self-explaining during fraction intervention. Journal of Educational Psychology, 108(4), 493.

Gagić, Z. Z., Skuban, S. J., Radulović, B. N., Stojanović, M. M., & Gajić, O. (2019). The implementation of mind maps in teaching physics: educational efficiency and students' involvement. Journal of Baltic Science Education, 18(1), 117-131.

Gerjets, P., Scheiter, K., & Schuh, J. (2008). Information comparisons in example-based hypermedia environments: Supporting learners with processing prompts and an interactive comparison tool. Educational Technology Research and Development, 56(1), 73-92.

Marqués, J. G., & Pelta, C. (2017). Concept maps and simulations in a computer system for learning Psychology. European Journal of education and Psychology, 10(1), 33-39.

Goossens, N. A., Camp, G., Verkoeijen, P. P., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. Journal of Applied Research in Memory and Cognition, 3(3), 177-182.

Griffin, C. C., & Jitendra, A. K. (2009). Word problem-solving instruction in inclusive third-grade mathematics classrooms. The Journal of Educational Research, 102(3), 187-202.

Guo, J. P., Yang, L. Y., & Ding, Y. (2014). Effects of example variability and prior knowledge in how students learn to solve equations. European journal of psychology of education, 29(1), 21-42.

Hadley, E. B., Dickinson, D. K., Hirsh-Pasek, K., & Golinkoff, R. M. (2019). Building semantic networks: The impact of a vocabulary intervention on preschoolers' depth of word knowledge. Reading Research Quarterly, 54(1), 41-61.

Herrmann-Abell, C. F., Koppal, M., & Roseman, J. E. (2016). Toward high school biology: Helping middle school students understand chemical reactions and conservation of mass in nonliving and living systems. CBE—Life Sciences Education, 15(4), ar74.

Hsieh, S. W., Ho, S. C., Wu, M. P., & Ni, C. Y. (2016). The Effects of concept map-oriented gesture-based teaching system on learners' learning performance and cognitive load in earth science course. Eurasia Journal of Mathematics, Science and Technology Education, 12(3), 621-635.

Huk, T., & Ludwigs, S. (2009). Combining cognitive and affective support in order to promote learning. Learning and Instruction, 19(6), 495-505.

Hwang, G. J., Kuo, F. R., Chen, N. S., & Ho, H. J. (2014). Effects of an integrated concept mapping and web-based problem-solving approach on students' learning achievements, perceptions and cognitive loads. Computers & Education, 71, 77-86.

Hwang, G. J., Zou, D., & Lin, J. (2020). Effects of a multi-level concept mapping-based question-posing approach on students' ubiquitous learning performance and perceptions. Computers & Education, 149, 103815.

Jitendra, A. K., Griffin, C. C., McGoey, K., Gardill, M. C., Bhat, P., & Riley, T. (1998). Effects of mathematical word problem solving by students at risk or with mild disabilities. The Journal of Educational Research, 91(6), 345-355.

Jitendra, A. K., Star, J. R., Starosta, K., Leh, J. M., Sood, S., Caskie, G., ... & Mack, T. R. (2009). Improving seventh grade students' learning of ratio and proportion: The role of schema-based instruction. Contemporary Educational Psychology, 34(3), 250-264.

Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. Educational Technology Research and Development, 53(3), 83-93.

Kern, C. L., & Crippen, K. J. (2017). The effect of scaffolding strategies for inscriptions and argumentation in a science cyberlearning environment. Journal of Science Education and Technology, 26(1), 33-43.

King, A., & Rosenshine, B. (1993). Effects of guided cooperative questioning on children's knowledge construction. The Journal of Experimental Education, 61(2), 127-148.

Korur, F., Toker, S., & Eryılmaz, A. (2016). Effects of the integrated online advance organizer teaching materials on students' science achievement and attitude. Journal of Science Education and Technology, 25(4), 628-640.

Leahy, W., & Sweller, J. (2005). Interactions among the imagination, expertise reversal, and element interactivity effects. Journal of Experimental Psychology: Applied, 11(4), 266.

Mason, L., & Sorzio, P. (1996). Analogical reasoning in restructuring scientific knowledge. European Journal of Psychology of Education, 11(1), 3-23.

McMaster, K. L., van den Broek, P., Espin, C. A., Pinto, V., Janda, B., Lam, E., ... & van Boekel, M. (2015). Developing a reading comprehension intervention: Translating cognitive theory to educational practice. Contemporary Educational Psychology, 40, 28-40.

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based programme to enhance study efficacy in a high and low risk setting. European Journal of Cognitive Psychology, 19(4-5), 743-768.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. Psychonomic Bulletin & Review, 14(2), 225-229.

Mwangi, W., & Sweller, J. (1998). Learning to solve compare word problems: The effect of example format and generating self-explanations. Cognition and instruction, 16(2), 173-199.

Ngu, B. H., Low, R., & Sweller, J. (2002). Text editing in chemistry instruction. Instructional Science, 30(5), 379-402.

Ngu, B. H., Mit, E., Shahbodin, F., & Tuovinen, J. (2009). Chemistry problem solving instruction: a comparison of three computer-based formats for learning from hierarchical network problem representations. Instructional Science, 37(1), 21-42.

Nokes, T. J., Hausmann, R. G., VanLehn, K., & Gershman, S. (2011). Testing the instructional fit hypothesis: the case of self-explanation prompts. Instructional Science, 39(5), 645-666.

Nokes-Malach, T. J., VanLehn, K., Belenky, D. M., Lichtenstein, M., & Cox, G. (2013). Coordinating principles and examples through analogy and self-explanation. European Journal of Psychology of Education, 28(4), 1237-1263.

Norqvist, M. (2018). The effect of explanations on mathematical reasoning tasks. International Journal of Mathematical Education in Science and Technology, 49(1), 15-30.

Okebukola, P. A., & Jegede, O. J. (1988). Cognitive preference and learning mode as determinants of meaningful learning through concept mapping. Science Education, 72(4), 489-500.

Okebukola, P. A. (1990). Attaining meaningful learning of concepts in genetics and ecology: An examination of the potency of the concept-mapping technique. Journal of research in science teaching, 27(5), 493-504.

Risko, V. J., & Alvarez, M. C. (1986). An investigation of poor readers' use of a thematic strategy to comprehend text. Reading research quarterly, 298-316.

Roelle, J., & Berthold, K. (2016). Effects of comparing contrasting cases and inventing on learning from subsequent instructional explanations. Instructional Science, 44(2), 147-176.

Roelle, J., & Renkl, A. (2020). Does an option to review instructional explanations enhance example-based learning? It depends on learners' academic self-concept. Journal of Educational Psychology, 112(1), 131.

Scheiter, K., Schleinschok, K., & Ainsworth, S. (2017). Why sketching may aid learning from science texts: Contrasting sketching with written explanations. Topics in Cognitive Science, 9(4), 866-882.

Schlag, S., & Ploetzner, R. (2011). Supporting learning from illustrated texts: Conceptualizing and evaluating a learning strategy. Instructional Science, 39(6), 921-937.

Schmid, R. F., & Telaro, G. (1990). Concept mapping as an instructional strategy for high school biology. The Journal of Educational Research, 84(2), 78-85.

Siler, S. A., & Klahr, D. (2016). Effects of terminological concreteness on middle-school students' learning of experimental design. Journal of Educational Psychology, 108(4), 547.

Sterner, G., Wolff, U., & Helenius, O. (2020). Reasoning about representations: effects of an early math intervention. Scandinavian Journal of Educational Research, 64(5), 782-800.

Sun, C. T., Ye, S. H., & Wang, Y. J. (2015). Effects of commercial video games on cognitive elaboration of physical concepts. Computers & Education, 88, 169-181.

Wang, Z., & Adesope, O. (2017). Do focused self-explanation prompts overcome seductive details? A multimedia study. Journal of Educational Technology & Society, 20(4), 47-57.

Weinstein, C. E. (1982). Training students to use elaboration learning strategies. Contemporary Educational Psychology, 7(4), 301-311.

Willoughby, T., Porter, L., Belsito, L., & Yearsley, T. (1999). Use of elaboration strategies by students in grades two, four, and six. The Elementary School Journal, 99(3), 221-231.

Wong, R. M., Adesope, O. O., & Carbonneau, K. J. (2020). Process-and product-oriented worked examples and self-explanations to improve learning performance. Journal of STEM Education: Innovations and Research, 20(2).

Ziegler, E., Edelsbrunner, P. A., & Stern, E. (2018). The relative merits of explicit and implicit learning of contrasted algebra principles. Educational Psychology Review, 30(2), 531-558.

Appendix 9: managing cognitive load

Summary of risk of bias (rob) analysis

		Bias					
Strategy	***Study***	***Randomisation process***	***Deviations from intended intervention***	***Missing outcome data***	***Measurement of the outcome***	***Selection of reported results***	***Overall***
Worked examples	Booth *et al.* (2015a)	Some concerns	Low	Some concerns	Low	Some concerns	Some concerns
Worked examples	Booth *et al.* (2015b)	Low	Low	Low	Low	Some concerns	Some concerns
Worked examples	Heemsoth *et al.* (2014)	Low	Low	Low	Low	Some concerns	Some concerns
Worked examples	McLaren *et al.* (2015)	Low	Low	Low	Low	Some concerns	Some concerns
Scaffolds and Guidance	Becker *et al.* (2020)	Some concerns	Low	Low	Low	Some concerns	Some concerns
Scaffolds and Guidance	Wijekumar *et al.* (2014)	Low	Low	Low	Low	Some concerns	Some concerns
Collaborative problem solving	Kirschner *et al.* (2011)	Low	Low	Low	Low	Some concerns	Some concerns

Database references – worked examples for problem-solving

Short Reference	Focus	Full Reference
Bentley *et al.* (2017)	Effect of worked examples on proportional reasoning problems in mathematics	Bentley, B., & Yates, G. C. (2017). Facilitating proportional reasoning through worked examples: Two classroom-based experiments. Cogent Education, 4(1), 1297213.
*Booth *et al.* (2015a)	Effect of AlgebraByExample assignments on algebra test scores	Booth, J. L., Oyer, M. H., Paré-Blagoev, E. J., Elliot, A. J., Barbieri, C., Augustine, A., & Koedinger, K. R. (2015). Learning algebra by example in real-world classrooms. Journal of Research on Educational Effectiveness, 8(4), 530-551.
*Booth *et al.* (2015b)	Effect of incorrect worked examples on algebra test scores	Booth, J. L., Cooper, L. A., Donovan, M. S., Huyghe, A., Koedinger, K. R., & Paré-Blagoev, E. J. (2015). Design-based research within the constraints of practice: AlgebraByExample. Journal of Education for Students Placed at Risk (JESPAR), 20(1-2), 79-100.
Bokosmaty *et al.* (2015)	Effect of worked example guidance type on geometry problem-solving	Bokosmaty, S., Sweller, J., & Kalyuga, S. (2015). Learning geometry problem solving by studying worked examples: Effects of learner guidance and expertise. American Educational Research Journal, 52(2), 307-333.
Kyun *et al.* (2009)	Effects of worked example presentations on algebraic problem solving	Kyun, S. A., & Lee, H. (2009). The effects of worked examples in computer-based instruction: Focus on the presentation format of worked examples and prior knowledge of learners. Asia pacific education review, 10(4), 495-503.
Mevarech *et al.* (2003)	Effect of metacognitive training versus worked examples on mathematical reasoning	Mevarech, Z. R., & Kramarski, B. (2003). The effects of metacognitive training versus worked-out examples on students' mathematical reasoning. British Journal of Educational Psychology, 73(4), 449-471.
Mulder *et al.* (2014)	Effect of heuristic worked examples on inquiry-based learning in physics	Mulder, Y. G., Lazonder, A. W., & de Jong, T. (2014). Using heuristic worked examples to promote inquiry-based learning. Learning and instruction, 29, 56-64.

Reed *et al.* (2013)	Effect of worked examples and Cognitive Tutor on constructing equations	Reed, S. K., Corbett, A., Hoffman, B., Wagner, A., & MacLaren, B. (2013). Effect of worked examples and Cognitive Tutor training on constructing equations. Instructional science, 41(1), 1-24.
Retnowati *et al.* (2010)	Effects of collaborative learning and task complexity on mathematics performance	Retnowati, E., Ayres, P., & Sweller, J. (2010). Worked example effects in individual and group work settings. Educational Psychology, 30(3), 349-367.
Retnowati *et al.* (2017)	Effects of collaborative learning and instructional format on mathematics performance	Retnowati, E., Ayres, P., & Sweller, J. (2017). Can collaborative learning improve the effectiveness of worked examples in learning mathematics?. Journal of educational psychology, 109(5), 666.
Van Gog *et al.* (2011)	Effect of worked example format on learning electrical circuits	Van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. Contemporary Educational Psychology, 36(3), 212-218.
Van Gog *et al.* (2012)	Effect of worked example type on learning electrical circuits	Van Gog, T., & Kester, L. (2012). A test of the testing effect: acquiring problem-solving skills from worked examples. Cognitive Science, 36(8), 1532-1541.
Van Loon-Hillen *et al.* (2012)	Effect of worked examples on subtraction performance	van Loon-Hillen, N., Van Gog, T., & Brand-Gruwel, S. (2012). Effects of worked examples in a primary school mathematics curriculum. Interactive Learning Environments, 20(1), 89-99.
Wong *et al.* (2019)	Effects of worked example type and self-explanation type on geometry test scores	Wong, R. M., Adesope, O. O., & Carbonneau, K. J. (2020). Process-and product-oriented worked examples and self-explanations to improve learning performance. Journal of STEM Education: Innovations and Research, 20(2).
Youssef-Shalala *et al.* (2014) - Expt. 3 only	Effect of worked examples on geometry problem solving	Youssef-Shalala, A., Ayres, P., Schubert, C., & Sweller, J. (2014). Using a general problem-solving strategy to promote transfer. Journal of Experimental Psychology: Applied, 20(3), 215.

* High eligibility study, identified for in-depth analysis

Database references – incomplete or incorrect worked examples

Short Reference	Focus	Full Reference
Baars *et al.* (2013)	Effect of partially worked-out examples on biology problem-solving and students' judgements of learning	Baars, M., Visser, S., Van Gog, T., de Bruin, A., & Paas, F. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. Contemporary Educational Psychology, 38(4), 395-406.
Grosse *et al.* (2018)	Effects of copying correct and incorrect solutions on mathematical problem solving	Große, C. S. (2018). "Copying allowed–But be careful, errors included!"–Effects of copying correct and incorrect solutions on learning outcomes. Learning and Instruction, 58, 173-181. Haslam, C. Y., & Hamilton, R. J. (2010). Investigating the use of integrated instructions to reduce the cognitive load associated with doing practical work in secondary school science. International journal of science education, 32(13), 1715-1737.
*Heemsoth *et al.* (2014)	Effect of incorrect examples on learning fractions	Heemsoth, T., & Heinze, A. (2014). The impact of incorrect examples on learning fractions: A field experiment with 6th grade students. Instructional Science, 42(4), 639-657.
McCann *et al.* (2019)	Effects of generating incorrect examples on algebra test scores	McCann, N. F. (2019). Using Error Anticipation Exercises as an Instructional Intervention in the Algebra Classroom (Doctoral dissertation, Temple University. Libraries).
*McLaren *et al.* (2015)	Effects of learning with erroneous examples on learning decimals with a web-based tutor	McLaren, B. M., Adams, D. M., & Mayer, R. E. (2015). Delayed learning effects with erroneous examples: a study of learning decimals with a web-based tutor. International Journal of Artificial Intelligence in Education, 25(4), 520-542.
Ngu *et al.* (2002)	Effects of text editing on chemistry problem solving	Ngu, B. H., Low, R., & Sweller, J. (2002). Text editing in chemistry instruction. Instructional Science, 30(5), 379-402.
Yang *et al.* (2016)^	Effects of collaborative learning and erroneous examples on subtraction knowledge	Yang, Z. K., Wang, M., Cheng, H. N., Liu, S. Y., Liu, L., & Chan, T. W. (2016). The Effects of learning from correct and erroneous examples in individual and collaborative settings. The Asia-Pacific Education Researcher, 25(2), 219-227.

* High priority study, identified for in-depth analysis; ^ = study included for more than one strategy.

Database references – scaffolds, guidance, and schema-based instruction

Short Reference	Focus	Full Reference
Becker *et al.** (2020)	Effect of tablets providing measurement data and lowering cognitive load to support experimental learning in physics.	Becker, S., Klein, P., Gößling, A., & Kuhn, J. (2020). Using mobile devices to enhance inquiry-based learning processes. Learning and Instruction, 69, 101350.
De Corte *et al.* (2001)	Effect of schema-based text comprehension strategies on reading comprehension skill	De Corte, E., Verschaffel, L., & Van De Ven, A. (2001). Improving text comprehension strategies in upper primary school children: A design experiment. British Journal of Educational Psychology, 71(4), 531-559.
Fuchs *et al.* (2004)	Effects of schema-based transfer instruction on real-life mathematical problem-solving	Fuchs, L. S., Fuchs, D., Finelli, R., Courey, S. J., & Hamlett, C. L. (2004). Expanding schema-based transfer instruction to help third graders solve real-life mathematical problems. American Educational Research Journal, 41(2), 419-445.
Fuchs *et al.* (2006)	Effects of schema-based instruction type on real-life mathematical problem solving	Fuchs, L. S., Fuchs, D., Finelli, R., Courey, S. J., Hamlett, C. L., Sones, E. M., & Hope, S. K. (2006). Teaching third graders about real-life mathematical problem solving: A randomized controlled study. The Elementary School Journal, 106(4), 293-311.
Hawlitschek *et al.* (2017)	Effects of instruction type on history learning in an educational game	Hawlitschek, A., & Joeckel, S. (2017). Increasing the effectiveness of digital educational games: The effects of a learning instruction on students' learning, motivation and cognitive load. Computers in Human Behavior, 72, 79-86.
Jitendra *et al.* (2009)	Effect of schema-based instruction on mathematical problem-solving	Jitendra, A. K., Star, J. R., Starosta, K., Leh, J. M., Sood, S., Caskie, G., ... & Mack, T. R. (2009). Improving seventh grade students' learning of ratio and proportion: The role of schema-based instruction. Contemporary Educational Psychology, 34(3), 250-264.
Li *et al.* (2007)	Effect of using databases on problem-based learning in science	Li, R., & Liu, M. (2007). Understanding the effects of databases as cognitive tools in a problem-based multimedia learning environment. Journal of Interactive Learning Research, 18(3), 345-363.
Oksa *et al.* (2010)	Effect of explanatory notes on reading comprehension and cognitive load (Expts. 1 and 3 only)	Oksa, A., Kalyuga, S., & Chandler, P. (2010). Expertise reversal effect in using explanatory notes for readers of Shakespearean text. Instructional Science, 38(3), 217-236.
Olina *et al.* (2006)	Effect of problem format and presentation sequence on grammatical knowledge	Olina, Z., Reiser, R., Huang, X., Lim, J., & Park, S. (2006). Problem format and presentation sequence: Effects on learning and mental effort among US high school students. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 20(3), 299-309.
Pawley *et al.* (2005)	Effects of explicit instruction and prior knowledge on equation formation	Pawley, D., Ayres*, P., Cooper, M., & Sweller, J. (2005). Translating words into equations: A cognitive load theory approach. Educational Psychology, 25(1), 75-97.
Richey *et al.* (2013): Expt. 1 only	Effect of adding explanations to worked examples on physics problem solving	Richey, J. E., & Nokes-Malach, T. J. (2013). How much is too much? Learning and motivation effects of adding instructional explanations to worked examples. Learning and Instruction, 25, 104-124.
Roelle *et al.* (2015)*	Two experiments: (1) The effect of focused processing prompts versus general instructions and (2) reduced explanations with prompts versus full explanations on science knowledge.	Roelle, J., Lehmkuhl, N., Beyer, M. U., & Berthold, K. (2015). The role of specificity, targeted learning activities, and prior knowledge for the effects of relevance instructions. Journal of Educational Psychology, 107(3), 705.
Salden *et al.* 2009)	Effects of tutored problem solving vs. fixed faded worked examples on mathematics performance	Salden, R. J., Aleven, V. A., Renkl, A., & Schwonke, R. (2009). Worked examples and tutored problem solving: redundant or synergistic forms of support?. Topics in Cognitive Science, 1(1), 203-213.

Wijekumar *et al*. (2012)	Effect of intelligent tutoring on nonfiction reading comprehension	Wijekumar, K. K., Meyer, B. J., & Lei, P. (2012). Large-scale randomized controlled trial with 4th graders using intelligent tutoring of the structure strategy to improve nonfiction reading comprehension. Educational Technology Research and Development, 60(6), 987-1013.
*Wijekumar *et al*. (2014)	Effect of intelligent tutoring on reading comprehension of expository texts	Wijekumar, K., Meyer, B. J., Lei, P. W., Lin, Y. C., Johnson, L. A., Spielvogel, J. A., ... & Cook, M. (2014). Multisite randomized controlled trial examining intelligent tutoring of structure strategy for fifth-grade readers. Journal of Research on Educational Effectiveness, 7(4), 331-357.
Wijekumar *et al*. (2017)	Effect of intelligent tutoring on recall of expository texts	Wijekumar, K., Meyer, B. J., Lei, P., Cheng, W., Ji, X., & Joshi, R. M. (2017). Evidence of an intelligent tutoring system as a mindtool to promote strategic memory of expository texts and comprehension with children in grades 4 and 5. Journal of Educational Computing Research, 55(7), 1022-1048.

* High eligibility study, identified for in-depth analysis

Database references – collaborative problem-solving

Short Reference	Focus	Full Reference
Dhlamini *et al*. (2013)	Effects of collaborative learning on mathematics performance	Dhlamini, J., & Mogari, D. (2013). The Effect of a Group Approach on the Performance of High School Mathematics Learners. Pythagoras, 34(2), 198.
*Kirschner *et al*. (2011)	Effects of collaborative learning and instructional format on biology test scores	Kirschner, F., Paas, F., Kirschner, P. A., & Janssen, J. (2011). Differential effects of problem-solving demands on individual and collaborative learning outcomes. Learning and Instruction, 21(4), 587-599.
Kirschner *et al*. (2009)	The effects of individual versus group learning (in triads) on biology test performance.	Kirschner, F., Paas, F., & Kirschner, P. A. (2009). Individual and group-based learning from complex cognitive tasks: Effects on retention and transfer efficiency. Computers in Human Behavior, 25(2), 306-314.
Retnowati *et al*. (2010)	Effects of collaborative learning and task complexity on mathematics performance	Retnowati, E., Ayres, P., & Sweller, J. (2010). Worked example effects in individual and group work settings. Educational Psychology, 30(3), 349-367.
Retnowati *et al*. (2017)	Effects of collaborative learning and instructional format on mathematics performance	Retnowati, E., Ayres, P., & Sweller, J. (2017). Can collaborative learning improve the effectiveness of worked examples in learning mathematics?. Journal of educational psychology, 109(5), 666.
Retnowati *et al*. (2018)	Effects of collaborative learning and prerequisite knowledge on mathematics performance	Retnowati, E., Ayres, P., & Sweller, J. (2018). Collaborative learning effects when students have complete or incomplete knowledge. Applied Cognitive Psychology, 32(6), 681-692.
Yang *et al*. (2016)^	Effects of collaborative learning and erroneous examples on subtraction knowledge	Yang, Z. K., Wang, M., Cheng, H. N., Liu, S. Y., Liu, L., & Chan, T. W. (2016). The Effects of learning from correct and erroneous examples in individual and collaborative settings. The Asia-Pacific Education Researcher, 25(2), 219-227.
Zambrano *et al*. (2019)	the effects of prior collaborative experience and density/distribution of information amongst collaborative learners in maths	Zambrano, J., Kirschner, F., Sweller, J., & Kirschner, P. A. (2019). Effects of group experience and information distribution on collaborative learning. Instructional Science, 47(5), 531-550.
Zhang *et al*. (2011)	The effects of two collaborative learning strategies (Open-ended and Task-based) with an individualized learning strategy on individual ICT learning in a computer-based environment	Zhang, L., Ayres, P., & Chan, K. (2011). Examining different types of collaborative learning in a complex computer-based environment: A cognitive load approach. Computers in Human Behavior, 27(1), 94-98.

* High priority study, identified for in-depth analysis; ^ = study included for more than one strategy.

References – working memory training

Studies identified for discussion.

Dunning, D. L., Holmes, J., & Gathercole, S. E. (2013). Does working memory training lead to generalized improvements in children with low working memory? A randomized controlled trial. Developmental Science, 16(6), 915-925.

Hitchcock, C., & Westwell, M. S. (2017). A cluster-randomised, controlled trial of the impact of Cogmed working memory training on both academic performance and regulation of social, emotional and behavioural challenges. Journal of Child Psychology and Psychiatry, 58(2), 140-150.

Roberts, G., Quach, J., Spencer-Smith, M., Anderson, P. J., Gathercole, S., Gold, L., ... & Wake, M. (2016). Academic outcomes 2 years after working memory training for children with low working memory: a randomized clinical trial. JAMA pediatrics, 170(5), e154568-e154568.

Rode, C., Robson, R., Purviance, A., Geary, D. C., & Mayr, U. (2014). Is working memory training effective? A study in a school setting. PloS one, 9(8), e104796.

Wright, H., Dorsett, R., Anders, J., Buzzeo, J., Runge, J., & Sanders, M. (2019). Improving Working Memory: Evaluation report and executive summary. Education Endowment Foundation.

Other studies

Alloway, T. P., Bibile, V., & Lau, G. (2013). Computerized working memory training: Can it lead to gains in cognitive skills in students?. Computers in Human Behavior, 29(3), 632-638.

Colmar, S., Double, K., Davis, N., Sheldon, L., Phillips, N., Cheng, M., & Briddon, S. (2020). Memory Mates: An evaluation of a classroom-based, student-focused working memory intervention. Journal of Psychologists and Counsellors in Schools, 30(2), 159-171.

Dahlin, K. I. (2011). Effects of working memory training on reading in children with special needs. Reading and writing, 24(4), 479-491.

Dahlin, K. I. (2013). Working memory training and the effect on mathematical achievement in children with attention deficits and special needs. Journal of Education and Learning, 2(1), 118-133.

Gray, S. A., Chaban, P., Martinussen, R., Goldberg, R., Gotlieb, H., Kronitz, R., ... & Tannock, R. (2012). Effects of a computerized working memory training program on working memory, attention, and academics in adolescents with severe LD and comorbid ADHD: a randomized controlled trial. Journal of Child Psychology and Psychiatry, 53(12), 1277-1284.

Jones, J. S., Milton, F., Mostazir, M., & Adlam, A. R. (2020). The academic outcomes of working memory and metacognitive strategy training in children: A double-blind randomized controlled trial. Developmental science, 23(4), e12870.

Layes, S., Lalonde, R., Bouakkaz, Y., & Rebai, M. (2018). Effectiveness of working memory training among children with dyscalculia: evidence for transfer effects on mathematical achievement—a pilot study. Cognitive processing, 19(3), 375-385.

Nemmi, F., Helander, E., Helenius, O., Almeida, R., Hassler, M., Räsänen, P., & Klingberg, T. (2016). Behavior and neuroimaging at baseline predict individual response to combined mathematical and working memory training in children. Developmental Cognitive Neuroscience, 20, 43-51.

Peng, P., & Fuchs, D. (2017). A randomized control trial of working memory training with and without strategy instruction: Effects on young children's working memory and comprehension. Journal of learning disabilities, 50(1), 62-80.

St Clair-Thompson, H., Stevens, R., Hunt, A., & Bolder, E. (2010). Improving children's working memory and classroom performance. Educational Psychology, 30(2), 203-219.

Studer-Luethi, B., Bauer, C., & Perrig, W. J. (2016). Working memory training in children: Effectiveness depends on temperament. Memory & Cognition, 44(2), 171-186.

van der Donk, M., Hiemstra-Beernink, A. C., Tjeenk-Kalff, A., Van Der Leij, A., & Lindauer, R. (2015). Cognitive training for children with ADHD: a randomized controlled trial of cogmed working memory training and 'paying attention in class'. Frontiers in psychology, 6, 1081.

Van der Molen, M., Van Luit, J. E. H., Van der Molen, M. W., Klugkist, I., & Jongmans, M. J. (2010). Effectiveness of a computerised working memory training in adolescents with mild to borderline intellectual disabilities. Journal of Intellectual Disability Research, 54(5), 433-447.

Zhang, H., Chang, L., Chen, X., Ma, L., & Zhou, R. (2018). Working memory updating training improves mathematics performance in middle school students with learning difficulties. Frontiers in human neuroscience, 12, 154.

Database references – wider evidence in this area

Ai, J., Yang, J., Zhang, T., Si, J., & Liu, Y. (2017). The Effect of Central Executive Load on Fourth and Sixth Graders' Use of Arithmetic Strategies. Psychologica Belgica, 57(2), 154.

Alam, K., & uz Zaman, T. (2011). Controlling Information Load Through Pre-Lecture Assignments and Students' Achievement in Mathematics at Secondary Level. Pakistan Journal of Education, 28(2).

Allen, M., & Vallée-Tourangeau, F. (2016). Interactivity defuses the impact of mathematics anxiety in primary school children. International Journal of Science and Mathematics Education, 14(8), 1553-1566.

Ardaç, D., & Unal, S. (2008). Does the amount of on-screen text influence student learning from a multimedia-based instructional unit?. Instructional Science, 36(1), 75-88.

Ashman, G., Kalyuga, S., & Sweller, J. (2020). Problem-solving or explicit instruction: Which should go first when element interactivity is high?. Educational Psychology Review, 32(1), 229-247.

Azevedo, R., Cromley, J. G., Winters, F. I., Moos, D. C., & Greene, J. A. (2005). Adaptive human scaffolding facilitates adolescents' self-regulated learning with hypermedia. Instructional science, 33(5-6), 381-412.

Beege, M., Schneider, S., Nebel, S., & Rey, G. D. (2020). Does the effect of enthusiasm in a pedagogical Agent's voice depend on mental load in the Learner's working memory?. Computers in Human Behavior, 112, 106483.

Bobis, J., Sweller, J., & Cooper, M. (1993). Cognitive load effects in a primary-school geometry task. Learning and Instruction, 3(1), 1-21.

Bos, F. A., Terlouw, C., & Pilot, A. (2009). The effect of a pretest in an interactive, multimodal pretraining system for learning science concepts. Educational research and evaluation, 15(6), 571-590.

Bourdin, B., & Fayol, M. (2000). Is graphic activity cognitively costly? A developmental approach. Reading and Writing, 13(3), 183-196.

Broadbent, H. J., Osborne, T., Rea, M., Peng, A., Mareschal, D., & Kirkham, N. Z. (2018). Incidental category learning and cognitive load in a multisensory environment across childhood. Developmental psychology, 54(6), 1020.

Bulu, S. T., & Pedersen, S. (2010). Scaffolding middle school students' content knowledge and ill-structured problem solving in a problem-based hypermedia learning environment. Educational Technology Research and Development, 58(5), 507-529.

Carroll, W. M. (1994). Using worked examples as an instructional support in the algebra classroom. Journal of educational psychology, 86(3), 360.

Cerpa, N., Chandler, P., & Sweller, J. (1996). Some conditions under which integrated computer-based training software can facilitate learning. Journal of Educational Computing Research, 15(4), 345-367.

Chen, O., Kalyuga, S., & Sweller, J. (2015). The worked example effect, the generation effect, and element interactivity. Journal of Educational Psychology, 107(3), 689.

Chen, O., Kalyuga, S., & Sweller, J. (2016). Relations between the worked example and generation effects on immediate and delayed tests. Learning and Instruction, 45, 20-30.

Chen, O., Kalyuga, S., & Sweller, J. (2016). When instructional guidance is needed. The Educational and Developmental Psychologist, 33(2), 149-162.

Chen, O., Retnowati, E., & Kalyuga, S. (2020). Element interactivity as a factor influencing the effectiveness of worked example–problem solving and problem solving–worked example sequences. British Journal of Educational Psychology, 90, 210-223.

Chu, H. C. (2014). Potential negative effects of mobile learning on students' learning achievement and cognitive load—A format assessment perspective. Journal of Educational Technology & Society, 17(1), 332-344.

Clarke, T., Ayres, P., & Sweller, J. (2005). The impact of sequencing and prior knowledge on learning mathematics through spreadsheet applications. Educational technology research and development, 53(3), 15-24.

Collins, M. F. (2016). Supporting inferential thinking in preschoolers: Effects of discussion on children's story comprehension. Early Education and Development, 27(7), 932-956.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the (): Effects of disfluency on educational outcomes. Cognition, 118(1), 111-115.

Es-Sajjade, A., & Paas, F. (2020). Educational theories and computer game design: lessons from an experiment in elementary mathematics education. Educational Technology Research and Development, 68(5), 2685-2703.

Fartoukh, M., Chanquoy, L., & Piolat, A. (2012). Effects of emotion on writing processes in children. Written Communication, 29(4), 391-411.

Fisher, A. V., Godwin, K. E., & Seltman, H. (2014). Visual environment, attention allocation, and learning in young children: When too much of a good thing may be bad. Psychological science, 25(7), 1362-1370.
Fong, S. F., Lily, L. P. L., & Por, F. P. (2012). Reducing cognitive overload among students of different anxiety levels using segmented animation. Procedia-Social and Behavioral Sciences, 47, 1448-1456.
Garner, R., Gillingham, M. G., & White, C. S. (1989). Effects of'seductive details' on macroprocessing and microprocessing in adults and children. Cognition and instruction, 6(1), 41-57.
Gerjets, P., Scheiter, K., & Schuh, J. (2008). Information comparisons in example-based hypermedia environments: Supporting learners with processing prompts and an interactive comparison tool. Educational Technology Research and Development, 56(1), 73-92.
Ginns, P., Chandler, P., & Sweller, J. (2003). When imagining information is effective. Contemporary Educational Psychology, 28(2), 229-251.
Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., & Renkl, A. (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. Learning and Instruction, 39, 72-87.
Glogger-Frey, I., Gaus, K., & Renkl, A. (2017). Learning from direct instruction: Best prepared by several self-regulated or guided invention activities?. Learning and Instruction, 51, 26-35.
Gnambs, T., Appel, M., & Kaspar, K. (2015). The effect of the color red on encoding and retrieval of declarative knowledge. Learning and Individual Differences, 42, 90-96.
Hennah, N. (2019). A novel practical pedagogy for terminal assessment. Chemistry Education Research and Practice, 20(1), 95-106.
Hoogerheide, V., Loyens, S. M., & Van Gog, T. (2014). Comparing the effects of worked examples and modeling examples on learning. Computers in Human Behavior, 41, 80-91.
Huang, K., Chen, C. H., Wu, W. S., & Chen, W. Y. (2015). Interactivity of question prompts and feedback on secondary students' science knowledge acquisition and cognitive load. Journal of Educational Technology & Society, 18(4), 159-171.
Hwang, G. J., Kuo, F. R., Chen, N. S., & Ho, H. J. (2014). Effects of an integrated concept mapping and web-based problem-solving approach on students' learning achievements, perceptions and cognitive loads. Computers & Education, 71, 77-86.
Jarraya, M., Rekik, G., Belkhir, Y., Chtourou, H., Nikolaidis, P. T., Rosemann, T., & Knechtle, B. (2019). Which presentation speed is better for learning basketball tactical actions through video modeling examples? The influence of content complexity. Frontiers in psychology, 10, 2356.
Jimenez, S. R., & Saylor, M. M. (2017). Preschoolers' word learning and story comprehension during shared book reading. Cognitive Development, 44, 57-68.
Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. Journal of educational psychology, 96(3), 558.
Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. Educational Technology Research and Development, 53(3), 83-93.
Kant, J. M., Scheiter, K., & Oschatz, K. (2017). How to sequence video modeling examples and inquiry tasks to foster scientific reasoning. Learning and Instruction, 52, 46-58.
Kern, C. L., & Crippen, K. J. (2017). The effect of scaffolding strategies for inscriptions and argumentation in a science cyberlearning environment. Journal of Science Education and Technology, 26(1), 33-43.
Kester, L., Kirschner, P. A., & van Merriënboer, J. J. (2006). Just-in-time information presentation: Improving learning a troubleshooting skill. Contemporary Educational Psychology, 31(2), 167-185.
Khng, K. H. (2017). A better state-of-mind: deep breathing reduces state anxiety and enhances test performance through regulating test cognitions in children. Cognition and Emotion, 31(7), 1502-1510.
Könings, K. D., van Zundert, M., & van Merriënboer, J. J. (2019). Scaffolding peer-assessment skills: Risk of interference with learning domain-specific skills?. Learning and Instruction, 60, 85-94.
Lang, J. W., & Lang, J. (2010). Priming competence diminishes the link between cognitive test anxiety and test performance: Implications for the interpretation of test scores. Psychological Science, 21(6), 811-819.
Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. Educational Psychology Review, 27(2), 291-304.
Lee, C. H., & Kalyuga, S. (2011). Effectiveness of different pinyin presentation formats in learning Chinese characters: A cognitive load perspective. Language Learning, 61(4), 1099-1118.
Leung, M., Low, R., & Sweller, J. (1997). Learning from equations or words. Instructional Science, 25(1), 37-70.
Likourezos, V., & Kalyuga, S. (2017). Instruction-first and problem-solving-first approaches: alternative pathways to learning complex tasks. Instructional Science, 45(2), 195-219.
Liu, H. C., & Chuang, H. H. (2011). Investigation of the impact of two verbal instruction formats and prior knowledge on student learning in a simulation-based learning environment. Interactive Learning Environments, 19(4), 433-446.
Liu, T. C., Lin, Y. C., & Paas, F. (2013). Effects of cues and real objects on learning in a mobile device supported environment. British Journal of Educational Technology, 44(3), 386-399.

Lyberg-Åhlander, V., Holm, L., Kastberg, T., Haake, M., Brännström, K. J., & Sahlén, B. (2015). Are children with stronger cognitive capacity more or less disturbed by classroom noise and dysphonic teachers?. International journal of speech-language pathology, 17(6), 577-588.

MacNabb, C., Schmitt, L., Michlin, M., Harris, I., Thomas, L., Chittendon, D., ... & Dubinsky, J. M. (2006). Neuroscience in middle schools: a professional development and resource program that models inquiry-based strategies and engages teachers in classroom implementation. CBE—Life Sciences Education, 5(2), 144-157.

Passolunghi, M. C., De Vita, C., & Pellizzoni, S. (2020). Math anxiety and math achievement: The effects of emotional and math strategy training. Developmental science, 23(6), e12964.

Matlen, B. J., & Klahr, D. (2013). Sequential effects of high and low instructional guidance on children's acquisition of experimentation skills: Is it all in the timing?. Instructional Science, 41(3), 621-634.

Mavilidi, M. F., Hoogerheide, V., & Paas, F. (2014). A quick and easy strategy to reduce test anxiety and enhance test performance. Applied Cognitive Psychology, 28(5), 720-726.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. The Journal of the Learning Sciences, 15(2), 153-191.

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. Journal of educational psychology, 87(2), 319.

Mwangi, W., & Sweller, J. (1998). Learning to solve compare word problems: The effect of example format and generating self-explanations. Cognition and instruction, 16(2), 173-199.

Neubrand, C., Borzikowsky, C., & Harms, U. (2016). Adaptive prompts for learning Evolution with worked examples-Highlighting the students between the" novices" and the" experts" in a classroom. International journal of environmental and science education, 11(14), 6774-6795.

Ngu, B. H., Mit, E., Shahbodin, F., & Tuovinen, J. (2009). Chemistry problem solving instruction: a comparison of three computer-based formats for learning from hierarchical network problem representations. Instructional Science, 37(1), 21-42.

Ngu, B. H., Yeung, A. S., & Tobias, S. (2014). Cognitive load in percentage change problems: unitary, pictorial, and equation approaches to instruction. Instructional Science, 42(5), 685-713.

Ngu, B. H., Phan, H. P., Hong, K. S., & Usop, H. (2016). Reducing intrinsic cognitive load in percentage change problems: The equation approach. Learning and Individual Differences, 51, 81-90.

Ngu, B. H., & Phan, H. P. (2017). Will learning to solve one-step equations pose a challenge to 8th grade students?. International Journal of Mathematical Education in Science and Technology, 48(6), 876-894.

Ngu, B. H., Yeung, A. S., Phan, H. P., Hong, K. S., & Usop, H. (2018). Learning to solve challenging percentage-change problems: a cross-cultural study from a cognitive load perspective. Journal of Experimental Education, 86(3), 362-385.

Ngu, B. H., Phan, H. P., Sigauke, A. T., Maniam, V., & Usop, H. (2019). Cognitive Load on Learning One-Step Equations: A Cross-Cultural Study between Australia and Malaysia. Current Politics and Economics of South, Southeastern, and Central Asia, 28(1), 103-130.

Nokes-Malach, T. J., VanLehn, K., Belenky, D. M., Lichtenstein, M., & Cox, G. (2013). Coordinating principles and examples through analogy and self-explanation. European Journal of Psychology of Education, 28(4), 1237-1263.

Oliver, M., Venville, G., & Adey, P. (2012). Effects of a cognitive acceleration programme in a low socioeconomic high school in regional Australia. International Journal of Science Education, 34(9), 1393-1410.

Owen, E., & Sweller, J. (1985). What do students learn while solving mathematics problems?. Journal of educational psychology, 77(3), 272.

Owens, P., & Sweller, J. (2008). Cognitive load theory and music instruction. Educational Psychology, 28(1), 29-45.

Park, B., Moreno, R., Seufert, T., & Brünken, R. (2011). Does cognitive load moderate the seductive details effect? A multimedia study. Computers in Human Behavior, 27(1), 5-10.

Pillay, H. K. (1994). Cognitive load and mental rotation: structuring orthographic projection for learning and problem solving. Instructional Science, 22(2), 91-113.

Pol, H. J., Harskamp, E. G., Suhre, C. J., & Goedhart, M. J. (2009). How indirect supportive digital help during and after solving physics problems can improve problem-solving abilities. Computers & Education, 53(1), 34-50.

Purnell, K. N., Solman, R. T., & Sweller, J. (1991). The effects of technical illustrations on cognitive load. Instructional Science, 20(5-6), 443-462.

Reisslein, J., Johnson, A. M., & Reisslein, M. (2014). Color coding of circuit quantities in introductory circuit analysis instruction. IEEE Transactions on Education, 58(1), 7-14.

Rekik, G., Khacharem, A., Belkhir, Y., Bali, N., & Jarraya, M. (2019). The effect of visualization format and content complexity on acquisition of tactical actions in basketball. Learning and Motivation, 65, 10-19.

Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. The Journal of Experimental Education, 70(4), 293-315.

Richter, J., Scheiter, K., & Eitel, A. (2018). Signaling text–picture relations in multimedia learning: The influence of prior knowledge. Journal of Educational Psychology, 110(4), 544.

Salden, R. J., Aleven, V., Schwonke, R., & Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. Instructional Science, 38(3), 289-307.
Schneider, S., Nebel, S., Beege, M., & Rey, G. D. (2018). The autonomy-enhancing effects of choice on cognitive load, motivation and learning with digital media. Learning and Instruction, 58, 161-172.
Schneider, S., Häßler, A., Habermeyer, T., Beege, M., & Rey, G. D. (2019). The more human, the higher the performance? Examining the effects of anthropomorphism on learning with media. Journal of educational psychology, 111(1), 57.
Schrader, C., & Bastiaens, T. J. (2012). The influence of virtual presence: Effects on experienced cognitive load and learning outcomes in educational computer games. Computers in Human Behavior, 28(2), 648-658.
Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. Computers in human behavior, 25(2), 258-266.
Schwonke, R., Renkl, A., Salden, R., & Aleven, V. (2011). Effects of different ratios of worked solution steps and problem solving opportunities on cognitive load and learning outcomes. Computers in Human Behavior, 27(1), 58-62.
Scrimin, S., Mason, L., & Moscardino, U. (2014). School-related stress and cognitive performance: A mood-induction study. Contemporary Educational Psychology, 39(4), 359-368.
Scruggs, T. E., Mastropieri, M. A., Brigham, F. J., & Sullivan, G. S. (1992). Effects of mnemonic reconstructions on the spatial learning of adolescents with learning disabilities. Learning Disability Quarterly, 15(3), 154-162.
Siler, S. A., & Klahr, D. (2016). Effects of terminological concreteness on middle-school students' learning of experimental design. Journal of Educational Psychology, 108(4), 547.
Snoder, P. (2017). Improving English Learners' Productive Collocation Knowledge: The Effects of Involvement Load, Spacing, and Intentionality. TESL Canada Journal, 34(3), 140-164.
Song, D. (2016). Expertise reversal effect and sequencing of learning tasks in online English as a second language learning environment. Interactive Learning Environments, 24(3), 423-437.
Spanjers, I. A., Wouters, P., Van Gog, T., & Van Merrienboer, J. J. (2011). An expertise reversal effect of segmentation in learning from animated worked-out examples. Computers in Human Behavior, 27(1), 46-52.
Hsu, Y. (2020). Teaching geometrics to young learners using computer-based simulation: the interaction effect of guidance, in relation to representation and manipulation, with socio-cultural background. Interactive Learning Environments, 1-17.
Thierry, K. L., Bryant, H. L., Nobles, S. S., & Norris, K. S. (2016). Two-year impact of a mindfulness-based program on preschoolers' self-regulation and academic performance. Early Education and Development, 27(6), 805-821.
Tindall-Ford, S., Agostinho, S., Bokosmaty, S., Paas, F., & Chandler, P. (2015). Computer-based learning of geometry from integrated and split-attention worked examples: The power of self-management. Computer, 89, 99.
Van Gog, T., Paas, F., & Van Merriënboer, J. J. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. Learning and Instruction, 16(2), 154-164.
Van Gog, T., Paas, F., & Van Merriënboer, J. J. (2008). Effects of studying sequences of process-oriented and product-oriented worked examples on troubleshooting transfer efficiency. Learning and Instruction, 18(3), 211-222.
van Zundert, M. J., Sluijsmans, D. M., Könings, K. D., & van Merriënboer, J. J. (2012). The differential effects of task complexity on domain-specific and peer assessment skills. Educational Psychology, 32(1), 127-145.
Van Zundert, M. J., Könings, K. D., Sluijsmans, D. M. A., & Van Merriënboer, J. J. G. (2012). Teaching domain-specific skills before peer assessment skills is superior to teaching them simultaneously. Educational Studies, 38(5), 541-557.
Wang, C. Y. (2015). Scaffolding middle school students' construction of scientific explanations: Comparing a cognitive versus a metacognitive evaluation approach. International Journal of Science Education, 37(2), 237-271.
Wang, Z., & Adesope, O. (2017). Do focused self-explanation prompts overcome seductive details? A multimedia study. Journal of Educational Technology & Society, 20(4), 47-57.
Wang, C., Fang, T., & Miao, R. (2018). Learning performance and cognitive load in mobile learning: Impact of interaction complexity. Journal of Computer Assisted Learning, 34(6), 917-927.
Wang, C., Fang, T., & Gu, Y. (2020). Learning performance and behavioral patterns of online collaborative learning: Impact of cognitive load and affordances of different multimedia. Computers & Education, 143, 103683.
Ward, M., & Sweller, J. (1990). Structuring effective worked examples. Cognition and instruction, 7(1), 1-39.
Yeung, A. S., Jin, P., & Sweller, J. (1998). Cognitive load and learner expertise: Split-attention and redundancy effects in reading with explanatory notes. Contemporary educational psychology, 23(1), 1-21.
Yeung, A. S. (1999). Cognitive load and learner expertise: Split-attention and redundancy effects in reading comprehension tasks with vocabulary definitions. The Journal of Experimental Education, 67(3), 197-217.
Hsu, Y., Gao, Y., Liu, T. C., & Sweller, J. (2015). Interactions Between Levels of Instructional Detail and Expertise When Learning with Computer Simulations. Educational Technology & Society, 18(4), 113-127.
Yung, H. I., & Paas, F. (2015). Effects of cueing by a pedagogical agent in an instructional animation: A cognitive load approach. Journal of Educational Technology & Society, 18(3), 153-160.

Appendix 10: Cognitive Theory of Multimedia Learning (Dual Coding)

Summary of risk of bias (rob) analysis

		Bias					
Strategy	***Study***	***Randomisation process***	***Deviations from intended intervention***	***Missing outcome data***	***Measurement of the outcome***	***Selection of reported results***	***Overall***
Visual representation or illustration	Csikos *et al.* (2012)	Low	Low	Low	Low	Some concerns	Some concerns
Visual representation or illustration	Lindner *et al.* (2017)	Low	Low	Low	Low	Some concerns	Some concerns
Visual representation or illustration	Lindner *et al.* (2020)	Low	Low	Low	Low	Some concerns	Some concerns
Diagrams	Bergey *et al.* (2015)	Low	Low	Some concerns	Low	Some concerns	Some concerns
Diagrams	Coleman *et al.* (2018)	Low	Low	Some concerns	Low	Some concerns	Some concerns
Diagrams	Cromley *et al.* (2016) (study2)	Low	Low	Low	Low	Some concerns	Some concerns
Spatial cognition, visualisation, and simulation	Barner *et al.* (2016)	Some concerns	Low	Low	Low	Some concerns	Some concerns
Spatial cognition, visualisation, and simulation	Lowrie *et al.* (2019)	High	Low	Low	Low	Some concerns	High

Database references – visual representation or illustration

Short Reference	Focus	Full Reference
Acha *et al.* (2009)	Effect of verbal and/or visual annotations on vocabulary learning	Acha, J. (2009). The effectiveness of multimedia programmes in children's vocabulary learning. British Journal of Educational Technology, 40(1), 23-31.
Aldalalah and Fong (2010)	Effect of audio/image/text on music theory learning among students of different music intelligence levels	Aldalalah, O., & Fong, S. F. (2010). Effects of computer–based instructional designs among pupils of different music intelligence levels. International Journal of Social Sciences, 5(3), 168-176.
Ardasheva *et al.* (2018)	Effect of representation and glossary label visuals on science outcomes	Ardasheva, Y., Wang, Z., Roo, A. K., Adesope, O. O., & Morrison, J. A. (2018). Representation visuals' impacts on science interest and reading comprehension of adolescent English learners. The Journal of Educational Research, 111(5), 631-643.
Barbieri *et al.* (2019)^	Intervention using number lines and 'incorporating key principles from the science of learning'.	Barbieri, C. A., Rodrigues, J., Dyson, N., & Jordan, N. C. (2020). Improving fraction understanding in sixth graders with mathematics difficulties: Effects of a number line approach combined with cognitive learning strategies. Journal of Educational Psychology, 112(3), 628.
Berends *et al.* (2009)	Effect of illustration type on arithmetic performance	Berends, I. E., & van Lieshout, E. C. (2009). The effect of illustrations in arithmetic problem-solving: Effects of increased cognitive load. Learning and Instruction, 19(4), 345-353.

Chiu *et al*. (2016)^	Effect of representation on algebra concept learning	Chiu, T. K., & Churchill, D. (2016). Design of learning objects for concept learning: Effects of multimedia learning principles and an instructional approach. Interactive Learning Environments, 24(6), 1355-1370.
Chiu *et al*. (2017)	Effect of visual aids and learner expertise on higher order mathematics thinking	Chiu, T. K., & Mok, I. A. (2017). Learner expertise and mathematics different order thinking skills in multimedia learning. Computers & Education, 107, 147-164.
Cohen *et al*. (2012)	Effect of picture representation on science vocabulary learning	Cohen, M. T., & Johnson, H. L. (2012). Improving the acquisition and retention of science material by fifth grade students through the use of imagery interventions. Instructional Science, 40(6), 925-955.
*Csikos *et al*. (2012)	Effect of visual representations on mathematical word problem solving	Csíkos, C., Szitányi, J., & Kelemen, R. (2012). The effects of using drawings in developing young children's mathematical word problem solving: A design experiment with third-grade Hungarian students. Educational studies in mathematics, 81(1), 47-65.
Diana *et al*. (1997)	Effect of geographic maps on geography fact learning	Diana, E. M., & Webb, J. M. (1997). Using geographic maps in classrooms: The conjoint influence of individual differences and dual coding on learning facts. learning and Individual Differences, 9(3), 195-214.
Edens *et al*. (2001)	Effect of pictorial representation on concept learning in science	Edens, K. M., & Potter, E. F. (2001). Promoting conceptual understanding through pictorial representation. Studies in Art Education, 42(3), 214-233.
Gambrell *et al*. (1993)	Effect of mental imagery and illustrations on reading performance	Gambrell, L. B., & Jawitz, P. B. (1993). Mental imagery, text illustrations, and children's story comprehension and recall. Reading Research Quarterly, 265-276.
Gerjets *et al*. (2009)	Effect of representational format and learner control on maths performance in a hypermedia environment	Gerjets, P., Scheiter, K., Opfermann, M., Hesse, F. W., & Eysink, T. H. (2009). Learning with hypermedia: The influence of representational formats and different levels of learner control on performance and learning behavior. Computers in Human Behavior, 25(2), 360-370.
Homer *et al*. (2010)	Effect of iconic representations on chemistry learning	Homer, B. D., & Plass, J. L. (2010). Expertise reversal for iconic representations in science visualizations. Instructional Science, 38(3), 259-276.
Kiili *et al*. (2006)	Effect of student-generated illustrations on learning about the human immune system	Kiili, K. (2006). Towards a participatory multimedia learning model. Education and Information Technologies, 11(1), 21-32.
Kuo *et al*. (2004)	Effect of mnemonic representation on learning Chinese characters	Kuo, M. L. A., & Hooper, S. (2004). The effects of visual and verbal coding mnemonics on learning Chinese characters in computer-based instruction. Educational technology research and development, 52(3), 23-34.
Kutbay *et al*. (2020)	Effect of animation representation on learning electricity	Kutbay, E., & Akpinar, Y. (2020). Investigating Modality, Redundancy and Signaling Principles with Abstract and Concrete Representation. International Journal of Education in Mathematics, Science and Technology, 8(2), 131-145.
Leopold *et al*. (2015)	Effect of instruction representation on science learning	Leopold, C., Doerner, M., Leutner, D., & Dutke, S. (2015). Effects of strategy instructions on learning from text and pictures. Instructional Science, 43(3), 345-364.
Leutner *et al*. (2009)	Effect of drawing and imagery on learning science from text	Leutner, D., Leopold, C., & Sumfleth, E. (2009). Cognitive load and science text comprehension: Effects of drawing and mentally imagining text content. Computers in Human Behavior, 25(2), 284-289.
*Lindner *et al*. (2017)	Effects of representational pictures to testing material on maths performance	Lindner, M. A., Lüdtke, O., Grund, S., & Köller, O. (2017). The merits of representational pictures in educational assessment: Evidence for cognitive and motivational effects in a time-on-task analysis. Contemporary Educational Psychology, 51, 482-492.
*Lindner *et al*. (2020)	Effect of representational and decorative pictures on performance in maths and science	Lindner, M. A. (2020). Representational and decorative pictures in science and mathematics tests: Do they make a difference?. Learning and Instruction, 68, 101345.
Liu *et al*. (2020)	Effect of dual-coding based computer-assisted learning on vocabulary learning	Liu, X., Liu, C. H., & Li, Y. (2020). The Effects of Computer-Assisted Learning Based on Dual Coding Theory. Symmetry, 12(5), 701.
Mason *et al*. (2013)	Effect of concrete and abstract illustrations on science learning	Mason, L., Pluchino, P., Tornatora, M. C., & Ariasi, N. (2013). An eye-tracking study of learning from science text with concrete and abstract illustrations. The Journal of Experimental Education, 81(3), 356-384.
Moreno *et al*. (1999)	Effect of multimedia-supported metaphors on arithmetic concept learning	Moreno, R., & Mayer, R. E. (1999). Multimedia-supported metaphors for meaning making in mathematics. Cognition and instruction, 17(3), 215-248.

Moreno *et al*. (2011): Experiments 1 & 2 only	Effect of concrete and abstract visual representations on learning about electric circuits	Moreno, R., Ozogul, G., & Reisslein, M. (2011). Teaching with concrete and abstract visual representations: Effects on students' problem solving, problem representations, and learning perceptions. Journal of educational psychology, 103(1), 32.
Prangsma *et al*. (2008)	Effect of collaborative construction of representations on learning in history	Prangsma, M. E., Van Boxtel, C. A., & Kanselaar, G. (2008). Developing a 'big picture': Effects of collaborative construction of multimodal representations in history. Instructional Science, 36(2), 117-136.
Purnell *et al*. (1991)	Effect of technical illustrations on geography comprehension	Purnell, K. N., & Solman, R. T. (1991). The influence of technical illustrations on students' comprehension in geography. Reading Research Quarterly, 277-299.
Richter *et al*. (2018)	Effect of 'signalling' text and/or pictures and prior knowledge on chemistry learning	Richter, J., Scheiter, K., & Eitel, A. (2018). Signaling text–picture relations in multimedia learning: The influence of prior knowledge. Journal of Educational Psychology, 110(4), 544.
Schlag *et al*. (2011)	Effect of a strategy to support learning from illustrated texts (about honeybees)	Schlag, S., & Ploetzner, R. (2011). Supporting learning from illustrated texts: Conceptualizing and evaluating a learning strategy. Instructional Science, 39(6), 921-937.
Schneider *et al*. (2018): Expt. 3 only	Effect of decorative pictures on learning from texts	Schneider, S., Dyrna, J., Meier, L., Beege, M., & Rey, G. D. (2018). How affective charge and text–picture connectedness moderate the impact of decorative pictures on multimedia learning. Journal of Educational Psychology, 110(2), 233.
Schneider *et al*. (2019)^	Effect of anthropomorphism on learning with media in science	Schneider, S., Häßler, A., Habermeyer, T., Beege, M., & Rey, G. D. (2019). The more human, the higher the performance? Examining the effects of anthropomorphism on learning with media. Journal of educational psychology, 111(1), 57.
Starbek *et al*. (2010)	Effect of animation on genetics knowledge and comprehension	Starbek, P., Starčič Erjavec, M., & Peklaj, C. (2010). Teaching genetics with multimedia results in better acquisition of knowledge and improvement in comprehension. Journal of Computer Assisted Learning, 26(3), 214-224.
Swanson *et al*. (2015)	Effect of verbal and/or visual strategies on mathematics problem solving	Swanson, H. L., Lussier, C. M., & Orosco, M. J. (2015). Cognitive strategies, working memory, and growth in word problem solving in children with math difficulties. Journal of Learning Disabilities, 48(4), 339-358.
Urhahne *et al*. (2009): Study 2 only	Effect of 2D or 3D computer simulations on understanding chemical structures	Urhahne, D., Nick, S., & Schanze, S. (2009). The effect of three-dimensional simulations on the understanding of chemical structures and their properties. Research in science education, 39(4), 495-513.

* High priority study, identified for in-depth analysis; ^ = study included for more than one strategy.

Database references – overview of studies: diagrams

Short Reference	Focus	Full Reference
*Bergey *et al*. (2015)	Effects of diagrams versus text on spaced restudy on biology knowledge and comprehension	Bergey, B. W., Cromley, J. G., Kirchgessner, M. L., & Newcombe, N. S. (2015). Using diagrams versus text for spaced restudy: Effects on learning in 10th grade biology classes. British Journal of Educational Psychology, 85(1), 59-74.
Booth *et al*. (2012)	Effect of diagrams, stories and equations on algebra problem solving	Booth, J. L., & Koedinger, K. R. (2012). Are diagrams always helpful tools? Developmental and individual differences in the effect of presentation format on student problem solving. British Journal of Educational Psychology, 82(3), 492-511.
Butcher and Aleven (2013)^	Effect of rule-diagram mapping on geometry learning	Butcher, K. R., & Aleven, V. (2013). Using student interactions to foster rule–diagram mapping during problem solving in an intelligent tutoring system. Journal of Educational Psychology, 105(4), 988.
Carson *et al*. (2003)	Effect of diagrammatic and text-based instructions on chemistry learning	Carlson, R., Chandler, P., & Sweller, J. (2003). Learning and understanding science instructional material. Journal of educational psychology, 95(3), 629.
Chen *et al*. (2019)	Effect of diagramming and summarising on learning from scientific texts	Chen, O., Manalo, E., & She, Y. (2019). Examining the influence of expertise on the effectiveness of diagramming and summarising when studying scientific materials. Educational Studies, 45(1), 57-71.
Chu *et al*. (2017)	Effect of diagrams on algebra equation problem solving	Chu, J., Rittle-Johnson, B., & Fyfe, E. R. (2017). Diagrams benefit symbolic problem-solving. British Journal of Educational Psychology, 87(2), 273-287.

*Coleman *et al*. (2018)	Effect of diagram design on comprehension of science texts	Coleman, J. M., McTigue, E. M., & Dantzler, J. A. (2018). What makes a diagram easy or hard? The impact of diagram design on fourth-grade students' comprehension of science texts. The elementary school journal, 119(1), 122-151.
Cromley *et al*. (2013a)	Effect of teaching diagram comprehension on comprehension of biology diagrams	Cromley, J. G., Perez, T. C., Fitzhugh, S. L., Newcombe, N. S., Wills, T. W., & Tanaka, J. C. (2013). Improving students' diagram comprehension with classroom instruction. The Journal of Experimental Education, 81(4), 511-537.
Cromley *et al*. (2013b)	Effect of teaching diagram comprehension on comprehension of biology diagrams	Cromley, J. G., Bergey, B. W., Fitzhugh, S., Newcombe, N., Wills, T. W., Shipley, T. F., & Tanaka, J. C. (2013). Effects of three diagram instruction methods on transfer of diagram comprehension skills: The critical role of inference while learning. Learning and Instruction, 26, 45-58.
*Cromley *et al*. (2016) (Study 2)^	Effect of cognitive science informed curriculum including teaching diagram comprehension in biology (Study 2)	Cromley, J. G., Weisberg, S. M., Dai, T., Newcombe, N. S., Schunn, C. D., Massey, C., & Merlino, F. J. (2016). Improving middle school science learning using diagrammatic reasoning. Science Education, 100(6), 1184-1213.
Kolloffel *et al*. (2009)	Effect of representational format on maths learning from an interactive computer simulation	Kolloffel, B., Eysink, T. H., de Jong, T., & Wilhelm, P. (2009). The effects of representational format on learning combinatorics from an interactive computer simulation. Instructional Science, 37(6), 503-517.
Purnell *et al*. (1992)	Effect of technical illustrations on cognitive load and learning in geography	Purnell, K. N., Solman, R. T., & Sweller, J. (1991). The effects of technical illustrations on cognitive load. Instructional Science, 20(5-6), 443-462.
Reisslein *et al*. (2015)	Effect of colour-coded diagrams on learning about electrical circuits	Reisslein, J., Johnson, A. M., & Reisslein, M. (2014). Color coding of circuit quantities in introductory circuit analysis instruction. IEEE Transactions on Education, 58(1), 7-14.
Swanson *et al*. (2013)	Effect of visual and schematic cognitive strategies on mathematics problem solving of children at risk of maths difficulties	Swanson, H. L., Lussier, C., & Orosco, M. (2013). Effects of cognitive strategy interventions and cognitive moderators on word problem solving in children at risk for problem solving difficulties. Learning Disabilities Research & Practice, 28(4), 170-183.

* High priority study, identified for in-depth analysis; ^ = study included for more than one strategy.

Database references – spatial cognition, visualisation, and simulation

Short Reference	Focus	Full Reference
*Barner *et al*. (2016)	Effect of mental abacus instruction on maths achievement	Barner, D., Alvarez, G., Sullivan, J., Brooks, N., Srinivasan, M., & Frank, M. C. (2016). Learning mathematics in a visuospatial format: A randomized, controlled trial of mental abacus instruction. Child Development, 87(4), 1146-1158.
Barner *et al*. (2018)	Effect of mental abacus instruction on maths achievement	Barner, D., Athanasopoulou, A., Chu, J., Lewis, M., Marchand, E., Schneider, R., & Frank, M. (2018). A one-year classroom-randomized trial of mental abacus instruction for first-and second-grade students.
Bokosmaty *et al*. (2017)	Effect of manipulating shapes (using mouse movements) on geometry problem-solving	Bokosmaty, S., Mavilidi, M. F., & Paas, F. (2017). Making versus observing manipulations of geometric properties of triangles to learn geometry using dynamic geometry software. Computers & Education, 113, 313-326.
De Koning *et al*. (2017)	Effect of mental simulation on reading comprehension	de Koning, B. B., Bos, L. T., Wassenburg, S. I., & van der Schoot, M. (2017). Effects of a reading strategy training aimed at improving mental simulation in primary school children. Educational Psychology Review, 29(4), 869-889.
Gilligan *et al*. (2019)	Effect of spatial training and instruction type on maths performance	Gilligan, K. A., Thomas, M. S., & Farran, E. K. (2020). First demonstration of effective spatial training for near transfer to spatial performance and far transfer to a range of mathematics skills at 8 years. Developmental science, 23(4), e12909.
Hawes *et al*. (2017)	Effect of spatial visualisation training on maths performance	Hawes, Z., Moss, J., Caswell, B., Naqvi, S., & MacKinnon, S. (2017). Enhancing children's spatial and numerical skills through a dynamic spatial approach to early geometry instruction: Effects of a 32-week intervention. Cognition and Instruction, 35(3), 236-264.

*Lowrie *et al*. (2019)	Effect of spatial visualisation training on maths performance	Lowrie, T., Logan, T., & Hegarty, M. (2019). The influence of spatial visualization training on students' spatial reasoning and mathematics performance. Journal of Cognition and Development, 20(5), 729-751.

* High eligibility study identified for in-depth analysis

Database references – wider evidence in this area

Ainsworth, S., Bibby, P., & Wood, D. (2002). Examining the effects of different multiple representational systems in learning primary mathematics. The Journal of the Learning Sciences, 11(1), 25-61.

Al-Abbasi, D. (2012). The effects of modality and multimedia comprehension on the performance of students with varied multimedia comprehension abilities when exposed to high complexity, self-paced multimedia instructional materials. Journal of Educational Multimedia and Hypermedia, 21(3), 215-239.

Al-Balushi, K. A., & Al-Balushi, S. M. (2018). Effectiveness of Brain-Based Learning for Grade Eight Students' Direct and Postponed Retention in Science. International Journal of Instruction, 11(3), 525-538.

Almasseri, M., & AlHojailan, M. I. (2019). How flipped learning based on the cognitive theory of multimedia learning affects students' academic achievements. Journal of Computer Assisted Learning, 35(6), 769-781.

Andrä, C., Mathias, B., Schwager, A., Macedonia, M., & von Kriegstein, K. (2020). Learning foreign language vocabulary with gestures and pictures enhances vocabulary memory for several months post-learning in eight-year-old school children. Educational Psychology Review, 32(3), 815-850.

Jägerskog, A. S., Jönsson, F. U., Selander, S., & Jonsson, B. (2019). Multimedia learning trumps retrieval practice in psychology teaching. Scandinavian journal of psychology, 60(3), 222-230.

Ayres, P., Marcus, N., Chan, C., & Qian, N. (2009). Learning hand manipulative tasks: When instructional animations are superior to equivalent static representations. Computers in Human Behavior, 25(2), 348-353.

Barak, M., & Dori, Y. J. (2011). Science education in primary schools: Is an animation worth a thousand pictures?. Journal of Science Education and Technology, 20(5), 608-620.

Baranowska, K. (2020). Learning most with least effort: subtitles and cognitive load. ELT Journal, 74(2), 105-115.

Boucheix, J. M., & Guignard, H. (2005). What animated illustrations conditions can improve technical document comprehension in young students? Format, signaling and control of the presentation. European Journal of Psychology of Education, 20(4), 369-388.

Boucheix, J. M., & Forestier, C. (2017). Reducing the transience effect of animations does not (always) lead to better performance in children learning a complex hand procedure. Computers in Human Behavior, 69, 358-370.

Brenner, M. E., Mayer, R. E., Moseley, B., Brar, T., Durán, R., Reed, B. S., & Webb, D. (1997). Learning by understanding: The role of multiple representations in learning algebra. American Educational Research Journal, 34(4), 663-689.

Broadbent, H. J., Osborne, T., Rea, M., Peng, A., Mareschal, D., & Kirkham, N. Z. (2018). Incidental category learning and cognitive load in a multisensory environment across childhood. Developmental psychology, 54(6), 1020.

Bruce, C. D., & Hawes, Z. (2015). The role of 2D and 3D mental rotation in mathematics for young children: what is it? Why does it matter? And what can we do about it?. ZDM, 47(3), 331-343.

Chan, T. K., Wong, S. W., Wong, A. M. Y., & Leung, V. W. H. (2019). The influence of presentation format of story on narrative production in Chinese children learning English-as-a-second-language: A comparison between graphic novel, illustration book and text. Journal of psycholinguistic research, 48(1), 221-242.

Leahy, W., Chandler, P., & Sweller, J. (2003). When auditory presentations should and should not be a component of multimedia instruction. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 17(4), 401-418.

Chen, C. M., & Sun, Y. C. (2012). Assessing the effects of different multimedia materials on emotions and learning performance for visual and verbal style learners. Computers & Education, 59(4), 1273-1285.

Chien, Y. T., & Chang, C. Y. (2012). Comparison of different instructional multimedia designs for improving student science-process skill learning. Journal of Science Education and Technology, 21(1), 106-113.

Cohen, M. T., & Johnson, H. L. (2011). Improving the acquisition of novel vocabulary through the use of imagery interventions. Early Childhood Education Journal, 38(5), 357-366.

Conradty, C., & Bogner, F. X. (2016). Hypertext or textbook: effects on motivation and gain in knowledge. Education Sciences, 6(3), 29.

Crollen, V., Noël, M. P., Honoré, N., Degroote, V., & Collignon, O. (2020). Investigating the respective contribution of sensory modalities and spatial disposition in numerical training. Journal of experimental child psychology, 190, 104729.

Darch, C., & Eaves, R. C. (1986). Visual displays to increase comprehension of high school learning-disabled students. The Journal of Special Education, 20(3), 309-318.

Dervić, D., Nermin, Đ. A. P. O., Mešić, V., & Đokić, R. (2019). Cognitive load in multimedia learning: an example from teaching about lenses. Journal of Education in Science Environment and Health, 5(1), 102-118.

Dindar, M., Kabakçı Yurdakul, I., & Dönmez, F. İ. (2015). Measuring cognitive load in test items: static graphics versus animated graphics. Journal of Computer Assisted Learning, 31(2), 148-161.
Fong, S. F., Lily, L. P. L., & Por, F. P. (2012). Reducing cognitive overload among students of different anxiety levels using segmented animation. Procedia-Social and Behavioral Sciences, 47, 1448-1456.
Fong, S. F. (2013). Effects of Segmented Animated Graphics among Students of Different Spatial Ability Levels: A Cognitive Load Perspective. Turkish Online Journal of Educational Technology-TOJET, 12(2), 89-96.
Gibbs, S. (2003). Do pictures make a difference? A test of the hypothesis that performance in tests of phonological awareness is eased by the presence of pictures. Educational Psychology in Practice, 19(3), 219-228.
Gruhn, S., Segers, E., & Verhoeven, L. (2020). Moderating role of reading comprehension in children's word learning with context versus pictures. Journal of computer assisted learning, 36(1), 29-45.
Schär, S. G., & Kaiser, J. (2006). Revising (multi-) media learning principles by applying a differentiated knowledge concept. International journal of human-computer studies, 64(10), 1061-1070.
Harskamp, E. G., Mayer, R. E., & Suhre, C. (2007). Does the modality principle for multimedia learning apply to science classrooms?. Learning and Instruction, 17(5), 465-477.
Kim, D., & Gilman, D. A. (2008). Effects of text, audio, and graphic aids in multimedia instruction for vocabulary learning. Journal of educational technology & society, 11(3), 114-126.
Lai, A. F., Chen, C. H., & Lee, G. Y. (2019). An augmented reality-based learning approach to enhancing students' science reading performances from the perspective of the cognitive load theory. British Journal of Educational Technology, 50(1), 232-247.
Leahy, W., & Sweller, J. (2008). The imagination effect increases with an increased intrinsic cognitive load. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 22(2), 273-283.
Leahy, W., & Sweller, J. (2011). Cognitive load theory, modality of presentation and the transient information effect. Applied Cognitive Psychology, 25(6), 943-951.
Leahy, W., & Sweller, J. (2016). Cognitive load theory and the effects of transient information on the modality effect. Instructional Science, 44(1), 107-123.
Lee, H., & Mayer, R. E. (2015). Visual aids to learning in a second language: Adding redundant video to an audio lecture. Applied Cognitive Psychology, 29(3), 445-454.
Leikin, R., Leikin, M., Waisman, I., & Shaul, S. (2013). Effect of the presence of external representations on accuracy and reaction time in solving mathematical double-choice problems by students of different levels of instruction. International Journal of Science and Mathematics Education, 11(5), 1049-1066.
Leslie, K. C., Low, R., Jin, P., & Sweller, J. (2012). Redundancy and expertise reversal effects when using educational technology to learn primary school science. Educational technology research and development, 60(1), 1-13.
Lin, C. C., & Yu, Y. C. (2017). Effects of presentation modes on mobile-assisted vocabulary learning and cognitive load. Interactive Learning Environments, 25(4), 528-542.
Martin, S. (2012). Does instructional format really matter? Cognitive load theory, multimedia and teaching English Literature. Educational Research and Evaluation, 18(2), 125-152.
Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents?. Cognition and instruction, 19(2), 177-213.
Moreno, R., & Durán, R. (2004). Do multiple representations need explanations? The role of verbal guidance and individual differences in multimedia mathematics learning. Journal of educational psychology, 96(3), 492.
Moses, A. M., Golos, D. B., & Bennett, C. M. (2015). An alternative approach to early literacy: The effects of ASL in educational media on literacy skills acquisition for hearing children. Early Childhood Education Journal, 43(6), 485-494.
Pujadas, G., & Muñoz, C. (2020). Examining adolescent EFL learners' TV viewing comprehension through captions and subtitles. Studies in Second Language Acquisition, 42(3), 551-575.
Rekik, G., Khacharem, A., Belkhir, Y., Bali, N., & Jarraya, M. (2019). The instructional benefits of dynamic visualizations in the acquisition of basketball tactical actions. Journal of Computer Assisted Learning, 35(1), 74-81.
Scheiter, K., Gerjets, P., & Schuh, J. (2010). The acquisition of problem-solving skills in mathematics: How animations can aid understanding of structural problem features and solution procedures. Instructional Science, 38(5), 487-502.
Scheiter, K., Schüler, A., Gerjets, P., Huk, T., & Hesse, F. W. (2014). Extending multimedia research: How do prerequisite knowledge and reading comprehension affect learning from text and pictures. Computers in Human Behavior, 31, 73-84.
Scheiter, K., Schleinschok, K., & Ainsworth, S. (2017). Why sketching may aid learning from science texts: Contrasting sketching with written explanations. Topics in Cognitive Science, 9(4), 866-882.
Scruggs, T. E., Mastropieri, M. A., Brigham, F. J., & Sullivan, G. S. (1992). Effects of mnemonic reconstructions on the spatial learning of adolescents with learning disabilities. Learning Disability Quarterly, 15(3), 154-162.
Segers, E., Verhoeven, L., & Hulstijn-Hendrikse, N. (2008). Cognitive processes in children's multimedia text learning. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 22(3), 375-387.

Spanjers, I. A., Wouters, P., Van Gog, T., & Van Merrienboer, J. J. (2011). An expertise reversal effect of segmentation in learning from animated worked-out examples. Computers in Human Behavior, 27(1), 46-52.
Stader, E. D. (1990). Children's Retrieval of Classroom Materials: A Test of Conjoint Retention.
Swanson, H. L., Orosco, M. J., & Lussier, C. M. (2014). The effects of mathematics strategy instruction for children with serious problem-solving difficulties. Exceptional Children, 80(2), 149-168.
Takacs, Z. K., & Bus, A. G. (2018). How pictures in picture storybooks support young children's story comprehension: An eye-tracking experiment. Journal of Experimental Child Psychology, 174, 1-12.
Taura, H. (1998). Bilingual dual coding in Japanese returnee students. Language Culture and Curriculum, 11(1), 47-70.
Tindall-Ford, S., & Sweller, J. (2006). Altering the modality of instructions to facilitate imagination: Interactions between the modality and imagination effects. Instructional science, 34(4), 343-365.
Türk, E., & Erçetin, G. (2014). Effects of interactive versus simultaneous display of multimedia glosses on L2 reading comprehension and incidental vocabulary learning. Computer Assisted Language Learning, 27(1), 1-25.
van Lieshout, E. C., & Xenidou-Dervou, I. (2018). Pictorial representations of simple arithmetic problems are not always helpful: a cognitive load perspective. Educational Studies in Mathematics, 98(1), 39-55.
Walma Van Der Molen, J. H., & Van Der Voort, T. H. (2000). The impact of television, print, and audio on children's recall of the news. A study of three alternative explanations for the dual-coding hypothesis. Human Communication Research, 26(1), 3-26.
Wang, C., Fang, T., & Gu, Y. (2020). Learning performance and behavioral patterns of online collaborative learning: Impact of cognitive load and affordances of different multimedia. Computers & Education, 143, 103683.
Weng, C., Otanga, S., Weng, A., & Cox, J. (2018). Effects of interactivity in E-textbooks on 7th graders science learning and cognitive load. Computers & Education, 120, 172-184.
Weng, C., Otanga, S., Christianto, S. M., & Chu, R. J. C. (2020). Enhancing students' biology learning by using augmented reality as a learning supplement. Journal of Educational Computing Research, 58(4), 747-770.
Witteman, M. J., & Segers, E. (2010). The modality effect tested in children in a user-paced multimedia environment. Journal of Computer Assisted Learning, 26(2), 132-142.
Wong, A., Leahy, W., Marcus, N., & Sweller, J. (2012). Cognitive load theory, the transient information effect and e-learning. Learning and instruction, 22(6), 449-457.
Wouters, P., Paas, F., & van Merriënboer, J. J. (2009). Observational learning from animated models: Effects of modality and reflection on transfer. Contemporary Educational Psychology, 34(1), 1-8.
Wouters, P., Paas, F., & Van Merriënboer, J. J. (2010). Observational learning from animated models: effects of studying–practicing alternation and illusion of control on transfer. Instructional Science, 38(1), 89-104.
Yang, C., Jen, C. H., Chang, C. Y., & Yeh, T. K. (2018). Comparison of animation and static-picture based instruction: Effects on performance and cognitive load for learning genetics. Journal of Educational Technology & Society, 21(4), 1-11.
Yung, H. I., & Paas, F. (2015). Effects of computer-based visual representation on mathematics learning and cognitive load.

Appendix 11: Embodied and Spatial Cognition

Summary of risk of bias (rob) analysis

		Bias					
Strategy	***Study***	***Randomisation process***	***Deviations from intended intervention***	***Missing outcome data***	***Measurement of the outcome***	***Selection of reported results***	***Overall***
Embodied cognition	Margolin *et al.* (2020)	Low	Low	Some concerns	Low	Some concerns	Some concerns

Database references – embodied cognition

Short Reference	Focus	Full Reference
Badinlou *et al.* (2018)	Effect of enactment cues on recall of action phrases	Badinlou, F., Kormi-Nouri, R., & Knopf, M. (2018). Action memory and knowledge-based cuing in school-aged children: The effect of object presentation and semantic integration. Acta psychologica, 186, 118-125.
Cook *et al.* (2013)	Effect of gestures on mathematical equivalence knowledge	Cook, S. W., Duffy, R. G., & Fenn, K. M. (2013). Consolidation and transfer of learning after observing hand gesture. Child development, 84(6), 1863-1871.
Corcoran *et al.* (2018)	Effect of embodied cognition (Mark DeGarmo Dance program) on reading achievement	Corcoran, R. P. (2018). An embodied cognition approach to enhancing reading achievement in New York City public schools: Promising evidence. Teaching and teacher education, 71, 78-85.
Ginns *et al.* (2016)	Effect of tracing worked examples on maths test scores	Ginns, P., Hu, F. T., Byrne, E., & Bobis, J. (2016). Learning by tracing worked examples. Applied Cognitive Psychology, 30(2), 160-169.
Hsaio *et al.* (2018)	Effect of gesture on plant knowledge and motor skills	Hsiao, H. S., Chen, J. C., Lin, C. Y., & Chen, W. N. (2018). The influence of a gesture-based learning approach on preschoolers' learning performance, motor skills, and motion behaviors. Interactive Learning Environments, 26(7), 869-881.
Hu *et al.* (2014)	Effect of tracing worked examples on geometry learning	Hu, F. T., Ginns, P., & Bobis, J. (2014). Does Tracing Worked Examples Enhance Geometry Learning?. Australian Journal of Educational & Developmental Psychology, 14, 45-49.
Kaschak *et al.* (2017)	Effect of gesture ('enacted reading') on abstract text comprehension	Kaschak, M. P., Connor, C. M., & Dombek, J. L. (2017). Enacted reading comprehension: Using bodily movement to aid the comprehension of abstract text content. PloS one, 12(1), e0169711.
Kosmas *et al.* (2019)	Effect of embodied learning on expressive vocabulary	Kosmas, P., Ioannou, A., & Zaphiris, P. (2019). Implementing embodied learning in the classroom: effects on children's memory and language skills. Educational Media International, 56(1), 59-74.
Kosmas *et al.* (2020)	Effect of embodied learning on expressive vocabulary	Kosmas, P., & Zaphiris, P. (2020). Words in action: investigating students' language acquisition and emotional performance through embodied learning. Innovation in Language Learning and Teaching, 14(4), 317-332.
*Margolin *et al.* (2020)	Effects of a play-based middle school physics program on physics knowledge	Margolin, J., Ba, H., Friedman, L. B., Swanlund, A., Dhillon, S., & Liu, F. (2020). Examining the impact of a play-based middle school physics program. Journal of Research on Technology in Education, 1-15.
Mavilidi *et al.* (2015)	Effects of whole-body movements (exercise) and part-body movements (gesture) on foreign language vocabulary performance	Mavilidi, M. F., Okely, A. D., Chandler, P., Cliff, D. P., & Paas, F. (2015). Effects of integrated physical exercises and gestures on preschool children's foreign language vocabulary learning. Educational psychology review, 27(3), 413-426.

Ruiter *et al*. (2015)	Effect of body movement on number knowledge	Ruiter, M., Loyens, S., & Paas, F. (2015). Watch your step children! Learning two-digit numbers through mirror-based observation of self-initiated body movements. Educational Psychology Review, 27(3), 457-474.
Schmidt *et al*. (2019)	Effect of embodied learning on foreign language vocabulary learning	Schmidt, M., Benzing, V., Wallman-Jones, A., Mavilidi, M. F., Lubans, D. R., & Paas, F. (2019). Embodied learning in the classroom: Effects on primary school children's attention and foreign language vocabulary learning. Psychology of sport and exercise, 43, 45-54.
Tang *et al*. (2019)	Effect of tracing on knowledge of the water cycle	Tang, M., Ginns, P., & Jacobson, M. J. (2019). Tracing enhances recall and transfer of knowledge of the water cycle. Educational Psychology Review, 31(2), 439-455.

* High eligibility study, identified for in-depth analysis

Database references – physical factors

Bunketorp Käll, L., Malmgren, H., Olsson, E., Lindén, T., & Nilsson, M. (2015). Effects of a curricular physical activity intervention on children's school performance, wellness, and brain development. Journal of School Health, 85(10), 704-713.

Dalziell, A., Booth, J. N., Boyle, J., & Mutrie, N. (2019). Better Movers and Thinkers: An evaluation of how a novel approach to teaching physical education can impact children's physical activity, coordination and cognition. British Educational Research Journal, 45(3), 576-591.

Özar, A. P. D. M. Does Classroom-based Physical Activity Influence Test Results?.

Donnelly, J. E., Greene, J. L., Gibson, C. A., Sullivan, D. K., Hansen, D. M., Hillman, C. H., ... & Washburn, R. A. (2013). Physical activity and academic achievement across the curriculum (A+ PAAC): rationale and design of a 3-year, cluster-randomized trial. BMC public health, 13(1), 1-8.

Egger, F., Benzing, V., Conzelmann, A., & Schmidt, M. (2019). Boost your brain, while having a break! The effects of long-term cognitively engaging physical activity breaks on children's executive functions and academic achievement. PloS one, 14(3), e0212482.

Fedewa, A. L., Ahn, S., Erwin, H., & Davis, M. C. (2015). A randomized controlled design investigating the effects of classroom-based physical activity on children's fluid intelligence and achievement. School Psychology International, 36(2), 135-153.

Gall, S., Adams, L., Joubert, N., Ludyga, S., Müller, I., Nqweniso, S., ... & Gerber, M. (2018). Effect of a 20-week physical activity intervention on selective attention and academic performance in children living in disadvantaged neighborhoods: A cluster randomized control trial. PloS one, 13(11), e0206908.

Have, M., Nielsen, J. H., Gejl, A. K., Ernst, M. T., Fredens, K., Støckel, J. T., ... & Kristensen, P. L. (2016). Rationale and design of a randomized controlled trial examining the effect of classroom-based physical activity on math achievement. BMC Public Health, 16(1), 1-11.

Howie, E. K., Schatz, J., & Pate, R. R. (2015). Acute effects of classroom exercise breaks on executive function and math performance: A dose–response study. Research Quarterly for Exercise and Sport, 86(3), 217-224.

Hraste, M., De Giorgio, A., Jelaska, P. M., Padulo, J., & Granić, I. (2018). When mathematics meets physical activity in the school-aged child: The effect of an integrated motor and cognitive approach to learning geometry. PLoS One, 13(8), e0196024.

Husain, F., Bartasevicius, V., Marshall, L., Chidley, S. & Forsyth, E. (2019) Fit to Study Evaluation Report. EEF. Available: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/fit-to-study/

Mavilidi, M. F., Okely, A. D., Chandler, P., & Paas, F. (2016). Infusing physical activities into the classroom: effects on preschool children's geography learning. Mind, Brain, and Education, 10(4), 256-263.

Mavilidi, M. F., Okely, A., Chandler, P., Domazet, S. L., & Paas, F. (2018). Immediate and delayed effects of integrating physical activity into preschool children's learning of numeracy skills. Journal of experimental child psychology, 166, 502-519.

Mavilidi, M. F., Ouwehand, K., Riley, N., Chandler, P., & Paas, F. (2020). Effects of an acute physical activity break on test anxiety and math test performance. International journal of environmental research and public health, 17(5), 1523.

Mavilidi, M. F., Drew, R., Morgan, P. J., Lubans, D. R., Schmidt, M., & Riley, N. (2020). Effects of different types of classroom physical activity breaks on children's on-task behaviour, academic achievement and cognition. Acta paediatrica, 109(1), 158-165.

Mullender-Wijnsma, M. J., Hartman, E., de Greeff, J. W., Doolaard, S., Bosker, R. J., & Visscher, C. (2016). Physically active math and language lessons improve academic achievement: a cluster randomized controlled trial. Pediatrics, 137(3).

Have, M., Nielsen, J. H., Ernst, M. T., Gejl, A. K., Fredens, K., Grøntved, A., & Kristensen, P. L. (2018). Classroom-based physical activity improves children's math achievement–A randomized controlled trial. PloS one, 13(12), e0208787.

Raney, M., Henriksen, A., & Minton, J. (2017). Impact of short duration health & science energizers in the elementary school classroom. Cogent Education, 4(1), 1399969.

Tandon, P. S., Klein, M., Saelens, B. E., Christakis, D. A., Marchese, A. J., & Lengua, L. (2018). Short term impact of physical activity vs. sedentary behavior on preschoolers' cognitive functions. Mental Health and Physical Activity, 15, 17-21.

Tarp, J., Domazet, S. L., Froberg, K., Hillman, C. H., Andersen, L. B., & Bugge, A. (2016). Effectiveness of a school-based physical activity intervention on cognitive performance in Danish adolescents: lcomotion—learning, cognition and motion–a cluster randomized controlled trial. PloS one, 11(6), e0158087.

Thompson, L. L. (2018). Take 5: An Analysis of the Effects of Classroom-based Physical Activity in Rural Middle Grades Mathematics Classes on Students' Grades, Test Scores, Cognitive Attitudes, and Academic Behaviors.

Tilp, M., Scharf, C., Payer, G., Presker, M., & Fink, A. (2020). Physical exercise during the morning school-break improves basic cognitive functions. Mind, Brain, and Education, 14(1), 24-31.

Tine, M. (2014). Acute aerobic exercise: an intervention for the selective visual attention and reading comprehension of low-income adolescents. Frontiers in Psychology, 5, 575.

Torbeyns, T., de Geus, B., Bailey, S., Decroix, L., Van Cutsem, J., De Pauw, K., & Meeusen, R. (2017). Bike desks in the classroom: Energy expenditure, physical health, cognitive performance, brain functioning, and academic performance. Journal of Physical Activity and Health, 14(6), 429-439.

Database references – wider evidence in this area

Adams, A. M., Glenberg, A. M., & Restrepo, M. A. (2019). Embodied reading in a transparent orthography. Learning and Instruction, 62, 27-36.

Andrä, C., Mathias, B., Schwager, A., Macedonia, M., & von Kriegstein, K. (2020). Learning foreign language vocabulary with gestures and pictures enhances vocabulary memory for several months post-learning in eight-year-old school children. Educational Psychology Review, 32(3), 815-850.

Bosse, M. L., Chaves, N., & Valdois, S. (2014). Lexical orthography acquisition: Is handwriting better than spelling aloud?. Frontiers in psychology, 5, 56.

Stegemann, K. C., & Grünke, M. (2014). Revisiting an old methodology for teaching counting, computation, and place value: the effectiveness of the finger calculation method for at-risk children. Learning Disabilities: A Contemporary Journal, 12(2), 191-213.

De Nooijer, J. A., Van Gog, T., Paas, F., & Zwaan, R. A. (2013). Effects of imitating gestures during encoding or during retrieval of novel verbs on children's test performance. Acta psychologica, 144(1), 173-179.

Margolin, J., Ba, H., Friedman, L. B., Swanlund, A., Dhillon, S., & Liu, F. (2020). Examining the impact of a play-based middle school physics program. Journal of Research on Technology in Education, 1-15.

Hsieh, S. W., Ho, S. C., Wu, M. P., & Ni, C. Y. (2016). The Effects of concept map-oriented gesture-based teaching system on learners' learning performance and cognitive load in earth science course. Eurasia Journal of Mathematics, Science and Technology Education, 12(3), 621-635.

Loeffler, J., Raab, M., & Cañal-Bruland, R. (2020). Let's do the time warp again–embodied learning of the concept of time in an applied school setting. Interactive Learning Environments, 1-10.

Mayer, C., Wallner, S., Budde-Spengler, N., Braunert, S., Arndt, P. A., & Kiefer, M. (2020). Literacy training of kindergarten children with pencil, keyboard or tablet stylus: The influence of the writing tool on reading and writing performance at the letter and word level. Frontiers in psychology, 10, 3054.

Novack, M. A., Congdon, E. L., Hemani-Lopez, N., & Goldin-Meadow, S. (2014). From action to abstraction: Using the hands to learn math. Psychological Science, 25(4), 903-910.

Singer, M. A., & Goldin-Meadow, S. (2005). Children learn when their teacher's gestures and speech differ. Psychological Science, 16(2), 85-89.

Toumpaniari, K., Loyens, S., Mavilidi, M. F., & Paas, F. (2015). Preschool children's foreign language vocabulary learning by embodying words through physical activity and gesturing. Educational Psychology Review, 27(3), 445-456.

Yeo, L. M., & Tzeng, Y. T. (2020). Cognitive effect of tracing gesture in the learning from mathematics worked examples. International Journal of Science and Mathematics Education, 18(4), 733-751.

Appendix 12: mixed strategy programmes

Summary of risk of bias (rob) analysis

	Bias					
Study	***Randomisation process***	***Deviations from intended intervention***	***Missing outcome data***	***Measurement of the outcome***	***Selection of reported results***	***Overall***
Cromley *et al.* (2016)^	Low	Low	Low	Low	Some concerns	Some concerns
Davenport *et al.* (2020)	Low	Some concerns	Some concerns	Low	Some concerns	Some concerns
Feddern *et al.* (2018)	Low	Low	Low	Low	Low	Low
Schunn *et al.* (2018)	Low	Low	Low	Low	Some concerns	Some concerns
Yang *et al.* (2020)	Some concerns	Some concerns	Low	Low	Some concerns	Some concerns

^ = study included for more than one strategy.

Database references – mixed strategy programmes

Study short reference	Focus	Full reference
Adey and Shayer (1993)	Lessons based on concrete activities, cognitive conflict, metacognition, schema development (bridging[47] of thinking strategies) in science.	Adey, P., & Shayer, M. (1993). An exploration of long-term far-transfer effects following an extended intervention program in the high school science curriculum. Cognition and instruction, 11(1), 1-29.
Barbieri *et al.* (2019)^	Intervention using number lines and 'incorporating key principles from the science of learning'.	Barbieri, C. A., Rodrigues, J., Dyson, N., & Jordan, N. C. (2020). Improving fraction understanding in sixth graders with mathematics difficulties: Effects of a number line approach combined with cognitive learning strategies. Journal of Educational Psychology, 112(3), 628.
*^Cromley *et al.* (2016)	Effect of cognitive science informed curriculum including teaching diagram comprehension in biology	Cromley, J. G., Weisberg, S. M., Dai, T., Newcombe, N. S., Schunn, C. D., Massey, C., & Merlino, F. J. (2016). Improving middle school science learning using diagrammatic reasoning. Science Education, 100(6), 1184-1213.
Davenport *et al.* (2020)*	Intervention to us CS concepts to revise a widely used middle school mathematics curriculum	Davenport, J. L., Kao, Y. S., Matlen, B. J., & Schneider, S. A. (2020). Cognition research in practice: engineering and evaluating a middle school math curriculum. The Journal of Experimental Education, 88(4), 516-535.
Yang *et al.* (2020)*	A comparison of training focused on cognitive science principles versus content knowledge in science	Yang, R., Porter, A. C., Massey, C. M., Merlino, J. F., & Desimone, L. M. (2020). Curriculum-based teacher professional development in middle school science: A comparison of training focused on cognitive science principles versus content knowledge. Journal of Research in Science Teaching, 57(4), 536-566.
Feddern *et al.* (2018)*	Testing the effectiveness of cognitive science-inspired biology revision software (spacing, interleaving, retrieval, visual cues) on biology test scores	Feddern, L., Belham, F. S., & Wilks, S. (2018). Retrieval, interleaving, spacing and visual cues as ways to improve independent learning outcomes at scale. Impact, Journal of the Chartered College of Teaching, 18, 19. Available: https://impact.chartered.college/article/feddern-retrieval-interleaving-spacing-visual-cues-independent-learning/

[47] i.e., strategies to generalise reasoning to promote transfer.

Schunn *et al*. (2018)* (also see implementation evidence in Desimone and Hill (2017))	Four principles of cognitive science were used to make systematic revisions in middle school science instructional modules from two kinds of curriculum	Schunn, C. D., Newcombe, N. S., Alfieri, L., Cromley, J. G., Massey, C., & Merlino, J. F. (2018). Using principles of cognitive science to improve science learning in middle school: What works when and for whom?. Applied cognitive psychology, 32(2), 225-240.

* High priority study, identified for in-depth analysis; ^ = study included for more than one strategy.

Database references – wider evidence in this area

Al-Balushi, K. A., & Al-Balushi, S. M. (2018). Effectiveness of Brain-Based Learning for Grade Eight Students' Direct and Postponed Retention in Science. International Journal of Instruction, 11(3), 525-538.

Desimone, L. M., & Hill, K. L. (2017). Inside the black box: Examining mediators and moderators of a middle school science intervention. Educational Evaluation and Policy Analysis, 39(3), 511-536.

Marqués, J. G., & Pelta, C. (2017). Concept maps and simulations in a computer system for learning Psychology. European Journal of education and Psychology, 10(1), 33-39.

Hennah, N. (2019). A novel practical pedagogy for terminal assessment. Chemistry Education Research and Practice, 20(1), 95-106.

Herrmann-Abell, C. F., Koppal, M., & Roseman, J. E. (2016). Toward high school biology: Helping middle school students understand chemical reactions and conservation of mass in nonliving and living systems. CBE—Life Sciences Education, 15(4), ar74.

MacNabb, C., Schmitt, L., Michlin, M., Harris, I., Thomas, L., Chittendon, D., ... & Dubinsky, J. M. (2006). Neuroscience in middle schools: a professional development and resource program that models inquiry-based strategies and engages teachers in classroom implementation. CBE—Life Sciences Education, 5(2), 144-157.

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based programme to enhance study efficacy in a high and low risk setting. European Journal of Cognitive Psychology, 19(4-5), 743-768.

Oliver, M., Venville, G., & Adey, P. (2012). Effects of a cognitive acceleration programme in a low socioeconomic high school in regional Australia. International Journal of Science Education, 34(9), 1393-1410.

Appendix 13: practice review

Practice review approach and methods

The overall objectives of this practice review are to understand:

1. What applications of cognitive science in the classroom are currently prominent in policy, guidance and practice? What do practitioners in England identify and recognize as common approaches based on cognitive science?
2. What form(s) do applications of cognitive science take when manifested in practice? How do cognitive science applications differ for different contexts, subjects and groups of students?

The practice review consists of three main parts:

1. **Literature review** of academic articles specifically discussing practice, reports, teaching frameworks and resources, and more popular-scientific texts and web-resources. The aims of the literature review were to identify how cognitive science is applied, understood and experienced by teachers in the classroom, including for different contexts, subjects and groups of students and to identify prominent policy and practice recommendations for practice in this area .

Literature for the review was located using two main strategies: 1) a general google search using the search terms: 'cognitive science' and 'classrooms.' This resulted in the identification of a large number of resources, including teaching frameworks, policy guidelines, practice reports and web-resources for teachers, 2) flagging relevant papers during our screening of papers for the systematic review. An outline of all included sources can be found in the appendix, where they are listed according to the topic they discuss.

The practical literature on cognitive science applications in education is vast, especially if you include teacher-generated resources, such as blogs and presentations, and the review is by no means exhaustive. We are aware that the internet is a major source of information for teachers, and therefore it could be argued that internet sources, such as twitter, might be included in the review. However, this was beyond the scope of the review, and would furthermore have to involve some decisions about what to include and where to draw the line. We make no claims to have covered all possible areas, but given the extent and scope of the resources reviewed, we argue that we present an overall picture of dominant practical approaches to cognitive science in education. This was supported by our interviews, which identified many of the same resources as the ones we had included. We were also able to identify some gaps, which we explored in further detail in the survey and qualitative interviews. By analysing the literature together with the survey data and qualitative interviews, we were thus able to give a comprehensive account of the diversity within each cognitive science strategy when applied to practice, and explore some of the many complexities with regards to subject and student diversity. Finally, as many of the resources provide examples of what teachers should be doing, rather than examples of what they are actually doing, the combination of the review, the survey and the qualitative interviews becomes invaluable to understand the link between theory and practice.

We did also review a number of popular books, although there were too many to do this as comprehensively as with reports and papers. We did consult policy guidance (such as the Ofsted summary of research) and resources from early career development programmes. Again, our treatment of these was exploratory and selective.

2. **Questionnaire survey of 808 practitioners**, mostly in England, to identify their experiences of using cognitive science in their classrooms , and their views of the applicability of cognitive science-inspired interventions for their particular subject and for different types of students . The survey was developed on the basis of the literature review and asked teachers a combination of quantitative and qualitative questions about their knowledge of different strategies, their use of them in their classrooms, any training they may have had and any issues they were encountering in relation to the adoption of cognitive science in education.

The questionnaire sample was a self-selected, non-representative sample of teachers. We distributed the questionnaire through organisational and personal contacts and via social media. We aimed to surface a range of perspectives and new ideas rather than seeking the results to be representative. Our sample is heavily skewed towards secondary teachers and more experienced teachers, as demonstrated by the sample characteristic results, below.

Table A13.1 – Education phases taught by teachers responding to the questionnaire

Education phase taught by survey respondent	%	Count
Nursery	10.0%	81
Primary	23.9%	193
Middle	5.3%	43
Secondary	50.5%	408
Further Education	8.4%	68
Other (state what)	1.9%	15
Total	100%	808

Table A13.2 – Years of teaching experience for teachers responding to the questionnaire

	Minimum	Maximum	Mean	Std Deviation	Variance	Count
How many years of experience do you have as an educator?	1.00	4.00	3.42	0.87	0.75	594

Table A13.3 – Years of teaching experience for teachers responding to the questionnaire

Years of experience as an educator	%	Count
0-2	5.2%	31
3-5	9.9%	59
6-10	22.7%	135
11 or more	62.1%	369
Total	100%	594

3. **Interviews with 13 practitioners**, discussing their experiences of cognitive science in more detail , particularly in relation to their views of how cognitive science applications may differ for different contexts, subjects and groups of students . Included in the interviews were also questions about where the participants saw the future of cognitive science in the classroom for themselves and the profession as a whole, and their views on what teachers need to know/would be interested in knowing more about . Participants for the interviews were selected from a large pool of survey respondents (200+) who had indicated that they were willing to be contacted for a follow-up interview. Based on the information they had given in the survey, we tried to select a diverse group of teachers in terms of gender, primary/secondary school, familiarity with cognitive science (low/medium/high), subjects and years of experience. The interviews were carried out in December 2020.

Table A13.4 – Characteristics of Interview respondents

Gender	***Phase***	***Reported level of familiarity with Cognitive Science* on Questionnaire**	***Other details provided of years of experience of teaching, role or subject***
Male N=6	Primary N= 4	Low/Medium N=2	Maths, 3-5 years' experience
			Class teacher, 6-10 years' experience
		Medium/High N=2	Head teacher, 11+ years' experience
			Maths, 3-5 years' experience
	Secondary N=2	Low/Medium N=1	Science, 11+ years' of experience
		Medium/High N=1	PE, 6-10 years' experience
Female N=7	Primary N=1	Low/Medium N=1	SENCO, 11+ years' of experience
	Secondary N=6	Low Medium/ N=3	English, 11+ years' of experience
			English, 6-10 years' experience
			Evidence lead, 6-10 years' experience
		Medium/High N=3	English, 11+ years' experience
			DT, 11+ years' experience
			Social science, 11 years' experience

Practice review bibliography – journal articles and chapters

Study short reference	Reference
Ahmed (2018)	Ahmed. F. (2018) First love letter to conflicting marriages: exploration of ethnically diverse students' developing understanding during their reading of Romeo and Juliet using schema theory, *English in Education*, 52:2, 105-119.
Alloway (2006)	Alloway, T. O. (2006) How does working memory work in the classroom? *Educational Research and Reviews* Vol. 1 (4), pp. 134-139
Clark and Mayer (2008)	Clark, R. C. and Mayer, R. E. (2008) Learning by viewing versus by doing: Evidence-based guidelines for Principled Learning environments, *Performance Improvemen*t, 47: 9, pp. 5-13.

Darling-Hammond *et al*. (2020)	Linda Darling-Hammond, Lisa Flook, Channa Cook-Harvey, Brigid Barron & David Osher (2020) Implications for educational practice of the science of learning and development, Applied Developmental Science, 24:2, 97-140, DOI: 10.1080/10888691.2018.1537791
Dunlosky and Rawson (2015)	Dunlosky, J. and Rawson, K. A. (2005) Practice Tests, Spaced Practice, and Successive Relearning: Tips for Classroom Use and for Guiding Students' Learning, Scholarship of Teaching and Learning in Psychology, 1, pp. 72-78
Dunlosky *et al*. (2013)	John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham (2013): Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology, Psychological Science in the Public Interest 14(1) 4–58
Fazio (2019)	Fazio, L. (2019) Retrieval practice opportunities in middle school mathematics teachers' oral questions, British Journal of Educational Psychology, 89, 653–669
Kelleher and Whitman (2018)	Kelleher, I. and Whitman, G. (2018) A Bridge No Longer Too Far: A Case Study of One School's Exploration of the Promise and Possibilities of Mind, Brain, and Education Science for the Future of Education, Mind, Brain and Education, vol 12 (4), pp. 224-230
Littrell-Baez *et al* (2015)	Megan K. Littrell-Baez , Angela Friend , Donna Caccamise , & Christine Okochi (2015) Using Retrieval Practice and Metacognitive Skills to Improve Content Learning, Journal of Adolescent & Adult Literacy 58(8), pp. 682-689
Miller and Endo (2004)	Miller, P.C. and Endo, H (2004) Understanding and Meeting The Needs of ESL Students, PHI DELTA KAPPAN, June 2004, pp. 786-791
Putnam and Roediger (2018)	Putnam, A. L. and Roediger, H. L. (2018) Education and Memory Seven Ways the Science of Memory Can Improve Classroom Learning, Chapter 6 in: Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Learning and Memory, John Wiley & Sons, Incorporated.
Ricommini and Morano 2019	Paul J. Riccomini and Stephanie Morano (2019) Guided Practice for Complex, Multistep Procedures in Algebra Scaffolding Through Worked Solutions, Teaching Exceptional Children, 15 (6) , pp. 445-454.
Jenkins (2018)	Jenkins, R. T (2018) Using educational neuroscience and psychology to teach science. Part 2: A case study review of 'The Brain-Targeted Teaching Model' and 'ResearchBased Strategies to Ignite Student Learning', School Science Review 100(371), pp. 66-75.
Van Gog and Rummel (2010)	van Gog, T. & Rummel, N. (2010) Example-Based Learning: Integrating Cognitive and Social-Cognitive Research Perspectives, Educ Psychol Rev (2010) 22:155–174
Vogel and Schwabe (2016)	Vogel, S. and Schwabe, L. (2016) Learning and memory under stress: implications for the classroom, npj Science of Learning 1, 16011; doi:10.1038/npjscilearn.2016.11;
Weinstein *et al* (2018)	Weinstein *et al*. Cognitive Research: Principles and Implications (2018) 3:2 DOI 10.1186/s41235-017-0087-y
Willis (2009)	Judy Willis (2009) What Brain Research Suggests for Teaching Reading Strategies, The Educational Forum, 73:4, 333-346, DOI: 10.1080/00131720903166861
Wittwer and Renkl (2010)	Wittwer, J and Renkl, A. (2010) How Effective are Instructional Explanations in Example-Based Learning? A Meta-Analytic Review, Educ Psychol Rev, 22:393–409
Yilmaz (2011)	Kaya Yilmaz (2011) The Cognitive Perspective on Learning: Its Theoretical Underpinnings and Implications for Classroom Practices, The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 84:5, 204-212, DOI: 10.1080/00098655.2011.568989

Practice review bibliography – reports

Study short reference	Reference
Coe *et al* (2020)	Coe, R. *et al* (2020) Great Teaching Toolkit - Evidence Review, Evidence Based Education and Cambridge Assessment International Education
Deans for Impact (2015)	Deans for Impact (2015). The Science of Learning. Austin, TX: Deans for Impact
CESE (2018)	CESE (2018) Cognitive load theory in practice - Examples for the classroom, Centre for Education Statistics and Evaluation, Sydney, Australia.
Howard-Jones (2014)	Howard-Jones, P. (2014) Neuroscience and Education: A Review of Educational Interventions and Approaches Informed by Neuroscience, EEF, University of Bristol.
Immordino - Yang *et al* (2018)	Mary Helen Immordino-Yang, Linda Darling-Hammond, Christina Krone (2018) The Brain Basis for Integrated Social, Emotional, and Academic Development How emotions and social relationships drive learning, The Aspen Institute National Commission on Social, Emotional, and Academic Development.
Hinton *et al* (2012)	Hinton, C., Fischer, K. W. and Glennon, C. (2012) Mind, Brain and Education, in: Teaching and Learning in the Era of the Common Core, Jobs for the Future.
Pashler *et al* (2007)	Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., and Metcalfe, J. (2007) Organizing Instruction and Study to Improve Student Learning (NCER 2007-2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ncer.ed.gov.
Rosenshine (2010)	Rosenshine, B. (2010) Principles of instruction, Educational Practice Series – 21, International Academy of Education.

Practice review bibliography – popular/magazine articles and opinion pieces

Study short reference	Reference
Booth (2018)	Booth, N. (2018) What does Research say about memory and what can we do to enhance long-term learning in the classroom, https://impact.chartered.college/article/booth-research-memory-learning-classroom/
Boser *et al* (2018)	Boser, U., McDaniels, A. and Benner, M. (2018) Using the Science of Learning to Redesign Schools, Center for American Progress.
Carey 2014	Carey, B. (2014) Brain Science in the Classroom, Education Week, nov. 4, 2014. https://www.edweek.org/teaching-learning/opinion-brain-science-in-the-classroom/2014/11?print=1
Caviglioli (2018)	Caviglioli, O. (2018) Six ways visuals help learning, https://impact.chartered.college/article/caviglioli-ways-visuals-helo-learning/
Dunlosky (2013)	Dynlosky (2013). Strengthening the Student Toolbox, study strategies to boost learning, American Educator, Fall 2013.
Firth (2018)	Firth, J. (2018) The Application of Spacing and Interleaving Approaches in the Classroom, https://impact.chartered.college/article/firth-spacing-interleaving-classroom/

Garner (2018)	Garner, S. (2018) Retrieval Practice in Use: Multiple-Choice testing in the primary classroom, https://impact.chartered.college/article/garner-retrieval-practice-multiple-choice-testing-primary-classroom/
Howard-Jones (2018)	Howard-Jones, P. (2018) Applying the Science of Learning in the Classroom, https://impact.chartered.college/article/howard-jones-applying-science-learning-classroom/
Immordino -Yang and Knecht (2020)	Immordino -Yang and Knecht (2020) Building Meaning Builds Teens' Brains, Educational leadership, May 2020, 77 (8), pp. 36-43
Shibli and West (2018)	Dominic Shibli and Rachel West (2018) Cognitive Load Theory and its application in the Classroom, https://impact.chartered.college/article/shibli-cognitive-load-theory-classroom/
Sumeracki and Weinstein (2018)	Sumeracki and Weinstein (2018) Optimising Learning using Retrieval Practice, https://impact.chartered.college/article/sumeracki-weinstein-optimising-learning-retrieval-practice/
Tomlinson and Sousa (2020)	Tomlinson, C. A. and Sousa, D. A. (2020) The Sciences of Learning, Educational Leadership, 77 (8), pp. 14-20)
WestEd 2014	WestEd. (2014). Merging cognitive science and curriculum to strengthen middle school math. R&D Alert, 15(1). San Francisco, CA:
Whitman (2018)	Whitman, G. (2018) Bridging the Gap between Mind, Brain and Educational Research and Practice; One school's replicable model, https://impact.chartered.college/category/building-a-science-of-learning/

Practice review bibliography – web-resources (incl. Online teacher resources)

Study short reference	Reference
Ambition institute	Ambition Institute – Teacher handbook: https://www.early-career-framework.education.gov.uk/ambition/wp-content/uploads/sites/3/2020/09/EarlyCareerTeachers_TextbookDigital_Teachers_v3-compressed.pdf
Educational Development Trust	Educational Development Trust, Early Career Professional Development Programme, https://www.educationdevelopmenttrust.com/ecf
Teach First	Teach First Early Careers Framework, https://www.teachfirst.org.uk/early-career-framework
UCL	UCL Early Career's framework Programme https://www.ucl.ac.uk/ioe/departments-and-centres/departments/learning-and-leadership/early-career-framework
The Learning scientists	https://www.learningscientists.org/
Boxer 2019	Boxer, A. (2019) 5 invaluable lessons from cognitive science, https://edu.rsc.org/feature/5-invaluable-lessons-from-cognitive-science/4010434.article
Hardiman (n.d.)	Hardinam, M. (n.d.) The Brain-Targeted Teaching Model: A Comprehensive Model for Instruction and Reform, http://braintargetedteaching.org/

Love to teach-blog (Kate Jones)	Examples of Dual Coding in the classroom https://lovetoteach87.com/2019/05/02/examples-of-dual-coding-in-the-classroom/
The effortful educator	https://theeffortfuleducator.com/2017/01/30/retrieval-practice-in-the-high-school-classroom/
Sumeracki (2016)	Dual coding - Can there be too much of a good thing? https://www.learningscientists.org/blog/2016/11/17-1

Table a13.1: list of relevant strategies and concepts by source

Relevant Strategies and Concepts	*Sources*	
Dual Coding		
▪ Still visuals vs. animation ▪ Accompanying visuals with narration or text ▪ Only including essential visuals ▪ Using computers to present text, visuals and sound in an integrated way ▪ Infographics ▪ Diagrams ▪ Graphic organiser ▪ Timelines ▪ Cartoon/Comic strips ▪ Notetaking using drawing ▪ visuals in the class material and comparing with text ▪ visuals and describe in your own words	Clark and Mayer (2008) Caviglioli (2018) Schmid (2008) Weinstein *et al* (2018) Love to teach blog Learning scientists Pashler *et al* (2007) WestEd 2014 Teach First The effortful educator Sumeracki (2016)	
Cognitive Load		
▪ Using visuals to reduce cognitive load (due to dual coding), but also making sure to avoid increasing cognitive load by including decorative, non-essential visuals. ▪ Strategies that work for 'novice learners' (limited, structured content and worked examples) vs. those that work for 'expert learners' (e.g. problem-solving) ▪ Breaking down material by using a 'part-whole' or a 'whole-part' approach ▪ Worked examples as a strategy to minimize cognitive load ▪ Cut out inessential material ▪ Cognitive load may be increased by other types of load (e.g., language load)	Ambition institute Deans for Impact (2015) Shibli and West (2018) Caviglioli (2018) Clark and Mayer (2008) Coe *et al* (2020) CESE (2018) Darling-Hammond *et al* (2020) Boxer (2019)	Ricommini and Morano 2019 Rosenshine (2010) Miller and Endo (2004) Wittwer and Renkl (2010) Van Gog and Rummel (2010) Teach First UCL Sumeracki (2016)
Spaced Learning		
▪ Within class spacing ▪ Across class spacing ▪ Combined with retrieval practice ▪ Spacing in the classroom ▪ Spacing in homework/study strategies ▪ 'Optimal' lag	Coe *et al* (2020) Deans for Impact (2015) Dunlovsky and Rawson (2015) Dunlovsky *et al* (2013) Dunlovsky (2013) Putnam and Roediger (2018) Weinstein *et al* (2018) Learning scientists	Howard-Jones (2014) Pashler *et al* (2007) Boser *et al* (2018) Carey 2014 Firth (2018) WestEd 2014 Whitman (2018) Teach First UCL
Retrieval Practice		
▪ Multiple choice tests ▪ Short answer tests ▪ Free recall practice tests ▪ Quizzes ▪ Production of concept maps	Coe *et al* (2020) Deans for Impact (2015) Garner (2018) Sumeracki and Weinstein (2018)	Rosenshine (2010) Learning scientists Ambition Institute Howard-Jones (2014a) Howard-Jones (2018)

▪ Timing of practice tests (delayed) ▪ Retrieval after close reading of text ▪ Semantic retrieval (of facts, procedures or events) ▪ Episodic retrieval (personal memories for experienced events). ▪ Retrieval with feedback ▪ The importance of retrieval 'success' ▪ Retrieval in class ▪ Retrieval as part of homework ▪ Weekly or monthly reviews ▪ Daily review (pre-questions or post-exposure)	Dunlovsky and Rawson (2015) Dunlovsky *et al* (2013) Dunlovsky (2013) Fazio (2018) Littrell-Baez *et al* (2015) Putnam and Roediger (2018) Weinstein *et al* (2018) Boxer (2019)	Pashler *et al* (2007) Sumeracki and Weinstein (2018) WestEd 2014 Whitman (2018) Ambition institute Teach First UCL Hardiman (n.d.) The effortful educator
Interleaving		
▪ Interleaving different types of problems ▪ Interleaving of inductive material ▪ Interleaving study and test opportunities	Deans for Impact (2015) Putnam and Roediger (2018) Weinstein *et al* (2018)	Learning scientists Pashler *et al* (2007) Carey 2014 Firth (2018)
Concrete examples		
	Weinstein *et al* (2018) Pashler *et al* (2007) Learning scientists	Boser *et al* (2018) Immordino-Yang and Knecht (2020)
Schema Theory/Prior Knowledge/pattern building/mental models		
▪ Connecting material to what students already know ▪ Making material relevant to the students frame of reference ▪ Making/developing categories	Coe *et al* (2020) Deans for Impact (2015) Ahmed (2018) Darling-Hammond *et al*. (2020) Ambition institute Yilmaz (2011) Willis (2009)	Booth (2018) Howard-Jones (2018) Ambition institute Educational Development Trust Teach First UCL
Worked examples		
▪ In-class ▪ As part of home work ▪ Correct solutions vs. erroneous solutions combined with feedback ▪ Pairing worked examples with problem-solving ▪ Completing or fading strategy	Ricommini and Morano 2019 Rosenshine (2010) Wittwer and Renkl (2010) Van Gog and Rummel (2010) WestEd 2014 UCL	
Working memory interventions or strategies		
▪ Breaking down instructions or activities ▪ Repeating and asking children to repeat instructions ▪ Memory aids ▪ Teaching children strategies to cope with working memory deficit ▪ Teacher-generated graphic organisers	Alloway (2006)	
Scaffolding		
▪ Teaching students within their Zone of Proximal Development. ▪ Provide assistance, reassurance and guidance to master a task beyond their existing ZPD. ▪ Teacher modelling examples or talking out load ▪ Providing students with cue cards, checklists or models of the completed task	Darling-Hammond *et al*. (2020) Rosenshine (2010)	
Elaboration		
▪ Involves adding features to existing knowledge ▪ Elaborative interrogation ▪ Asking 'why' questions while reading a text. ▪ Self-explain while learning	Putnam and Roediger (2018) Weinstein *et al* (2018) Learning scientists	

Self-explanation	
▪ Content specific or content neutral questions	Putnam and Roediger (2018)
Memory tools	
▪ Mnemonics: ▪ single use mnemonics vs. multiple use mnemonics ▪ Journey method. ▪ To provide an organisational structure to material that is unorganised ▪ To re-encode 'to be remembered' information into a format where it can be more easily remembered, often by using visual imaginary (Putnam and Roediger, p. 192)	Boser (2018) Putnam and Roediger (2018) Booth (2018)
Whole child	
▪ Social and Emotional experiences of the child ▪ Age-appropriate ▪ Fostering habits of mind (curiosity, empathy, awareness)	Darling-Hammond *et al* (2020) Immordino-Yang *et al* (2018) Immordino-Yang and Knecht (2020)
Growth mindset/Brain plasticity	
▪ Teaching students about a growth mindset ▪ Adopt flexible environment for teaching and learning ▪ Developing a classroom environment for achievement and hard work ▪ Active learning experiences	Tomlinson and Sousa (2020) Hinton (2012)
Broader programmes	
▪ Mind, Brain and Education ▪ Brain-targeted teaching	Keheller and Whitman (2018) Jenkins (2018) Hardiman (n.d.)
Emotional aspects	
▪ Negative impact of: ▪ Anxiety ▪ Stereotypes ▪ Stress ▪ Toxic stress ▪ Trauma ▪ Test anxiety ▪ Positive impact of: ▪ Excitement about learning ▪ Safety ▪ Belonging ▪ Mindfulness	Darling-Hammond *et al*. (2020) Miller and Endo (2004) Jenkins (2018) Weinstein *et al* (2018) Hardiman (n.d.) Hinton *et al* (2012) Vogel and Schwabe (2016) Howard-Jones (2014a) Tomlinson and Sousa (2020) Educational Development Trust
School/Classroom Environment	
▪ Cooperative rather than competitive classroom culture ▪ Emotionally supportive environment	Darling-Hammond *et al*. (2020) Dunlovsky *et al*. (2013) Jenkins (2018) Immodino-Yang *et al* (2018) Hinton *et al* (2012)