



## **Effective approaches to teaching mathematics in Key Stages 3 and 4**

Protocol for a systematic review and meta-analysis

February 2024 (*Updated: March 2025*)

Jeremy Hodgen, Rachel Marks, Eirini Geraniou, Nicola Bretscher, Laurie Jacques



The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus Trust (now part of Impetus - Private Equity Foundation) and received a founding £125m grant from the Department for Education.

Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

For more information about the EEF or this report please contact:



Education Endowment Foundation  
5th Floor, Millbank Tower  
21–24 Millbank  
SW1P 4QP



0207 802 1653



[info@eefoundation.org.uk](mailto:info@eefoundation.org.uk)



[www.educationendowmentfoundation.org.uk](http://www.educationendowmentfoundation.org.uk)



# Table of contents

<b>Background and review rationale .....</b>	<b>5</b>
Objectives.....	6
Research questions.....	6
Advisory group.....	8
<b>Methodology.....</b>	<b>9</b>
Logic of the review .....	9
Topic identification .....	10
Information management.....	10
Inclusion and exclusion criteria for the review.....	13
Search strategy for identification of studies .....	17
Phase 1: Full systematic search for RCTs and high-quality QED studies across all Research Questions and Topics .....	17
Phase 2: Systematic search for 'lower quality' QEDs and broader quantitative studies across selected Research Questions and Topics .....	24
Exemplary Reviews .....	25
Data extraction and management .....	25
Data synthesis .....	28
<b>Reporting .....</b>	<b>29</b>
Modular (RQ1) and RQ4 / RQ5 reporting .....	29
Narrative Reporting .....	29
Quality of evidence .....	30
Relevance of studies to mathematics teaching in England's secondary schools.....	31
Responding to a lack of quantitative evidence: Exemplary reviews .....	32
Key features of successful approaches (RQ2).....	32
<b>Registration, data protection and ethics .....</b>	<b>33</b>
<b>Team.....</b>	<b>33</b>
<b>Conflicts of interest .....</b>	<b>34</b>

<b>Timeline.....</b>	<b>34</b>
<b>Protocol Amendments.....</b>	<b>36</b>
<b>References .....</b>	<b>38</b>
<b>Appendix 1 – PISA Participants.....</b>	<b>42</b>
<b>Appendix 2 – Data extraction and coding tools .....</b>	<b>43</b>
Appendix 2a – EEF main data extraction tool (version 2) (EEF, 2022a) .....	43
Appendix 2b – EEF effect size data extraction tool (version 2) (EEF, 2022b) .....	46
Appendix 2c – Topic and intervention evaluation coding tool.....	50
Appendix 2d – Assessment of risk of bias coding tool.....	53
<b>Appendix 3 – Pseudo code for meta-analysis .....</b>	<b>54</b>
<b>Appendix 4 – Quality Assessment for Individual Studies .....</b>	<b>56</b>
<b>Appendix 5 – Relevance Assessment for Individual Studies .....</b>	<b>57</b>

## Background and review rationale

A good understanding of mathematics has significant benefits for individuals (Adkins & Noyes, 2012; Parsons & Bynner, 2005), whilst a mathematically skilled workforce is essential for a strong economy (Deloitte, 2012; Hodgen & Marks, 2013). Hence, it is no surprise that, around the world, raising attainment in mathematics is a key concern for both policy-makers and education practitioners (Schmidt et al., 2022). However, addressing this problem has proved to be difficult, particularly in secondary mathematics. In England, for example, despite more than two decades of reform initiatives, student performance in secondary mathematics has shown only very small improvements over time and lags well behind that of primary mathematics (Richardson et al., 2020). Low attainment in mathematics is particularly acute for disadvantaged students (Cooke et al., 2020; Jerrim et al., 2018). This lack of success is, in large part, due to a lack of robust and systematically reviewed evidence about effective approaches to raising mathematical attainment for secondary pupils. Indeed, practitioners and policy-makers often profoundly disagree about mathematics teaching method. See, for example, the controversy in England surrounding the recent OFSTED, 2021, Research Review: Gilmore et al., 2021.

Over the past two decades, there has been a plethora of systematic reviews, meta-analyses and secondary reviews of aspects of mathematics teaching and learning that have begun to address the question of how to teach mathematics. Indeed, two recent secondary meta-analyses (Hodgen et al., 2018, 2020a) identified more than 100 meta-analyses of relevance to upper primary and lower secondary mathematics. These reviews have in large part been made possible by two policy decisions in the US and England, respectively, that encouraged the use of randomised controlled trials (RCTs) and other rigorous experimental designs to evaluate educational interventions: the establishment of the Institute of Education Science in the US (2002) and the Education Endowment Foundation (EEF) in England (2011).

The existing corpus of systematic reviews in mathematics education includes meta-analyses and best-evidence reviews addressing effective approaches to specific approaches to aspects of mathematics teaching (e.g., Kyriacou & Issitt, 2008; Woodward et al., 2012), the use of resources and technology (e.g., Carbonneau et al., 2013; Young, 2017) and to the teaching of particular topics (e.g., Rakes et al., 2010; Siegler et al., 2010). Others have examined effective mathematics teaching for specific groups of students, such as students with low prior attainment (Baker et al., 2002), 'struggling' students (Fuchs et al., 2021), students with mathematical learning disabilities (e.g., Gersten et al., 2009) or young children (Simms et al., 2019). Yet, despite Seidel & Shavelson's (2007) finding that domain-specific components of teaching have a very much larger effect than generic components, few systematic reviews and meta-analyses have examined, and compared, the approaches to mathematics teaching for secondary students in general. Slavin et al.'s (2009) best-evidence synthesis examined effective programmes in middle and high school mathematics, but their study excluded interventions lasting less than 12 weeks and is now more than a decade out of date. Hodgen et al.'s (2018) more recent secondary meta-analysis addressed this to some extent, but was focused exclusively on synthesising the results of reviews and meta-analyses rather than primary studies and was restricted to lower secondary mathematics. Additionally, we are aware of only one high-quality meta-analysis that has examined effective approaches for teaching targeted at students with low socio-economic status (Dietrichson et al., 2017), which was focused on teaching in general rather than mathematics-specific approaches, although many of the included primary studies were conducted in the context of mathematics.

This study will address this gap by conducting a systematic review and meta-analysis of rigorous experimental studies of interventions focused on secondary mathematics. Our review will take the following approach. First, we will examine the evidence about the efficacy in raising mathematical attainment of different teaching interventions relevant to secondary mathematics (ages 11-16). In order to identify the intervention types, we will draw on three sources: two previous systematic reviews conducted by members of the review team (Hodgen et al., 2018, 2020a), a practice review that has been commissioned by the Education Endowment Foundation (EEF) in parallel to this review (Boylan, Forthcoming), and an analysis of relevant meta-analyses identified using a systematic search strategy. We will investigate whether there are differential effects for socio-economically disadvantaged pupils as well as for pupils with low prior attainment and for girls. Second, we will identify a set of pedagogic and implementation components, each common to several interventions (within and between topics). Then, through a comparison of all interventions, we will identify components and clusters of components that appear to be associated with effective interventions. Third, we will examine the specific relevance of interventions to mathematics classrooms at KS3 and KS4 in England. Our review will

synthesise the international research base and will therefore have international relevance. However, the review has been commissioned to inform the EEF's work in England and, hence, the relevance to English classrooms is important.

In this protocol, we follow the EEF template for evidence reviews (EEF, 2023) and the PRISMA guidance (Page et al., 2021) in pre-specifying our methods and approach.

## Objectives

The objective of this review is to synthesise the existing literature to identify effective interventions, approaches and strategies to teaching mathematics in Key Stages 3 and 4 in England, including the transitions KS2→KS3 and KS4→KS5 and within this to identify key research gaps that could influence EEF funding rounds in 2024 and beyond.

## Research questions

To achieve our objective, we will address the following research questions:

### ***[RQ1]. What is the evidence on the effectiveness of different approaches for teaching mathematics in Key Stages 3 and 4?***

For the purposes of this review, we follow Simms et al. (2019) in defining approaches, or interventions, as a clearly described change from, or difference to, existing, or usual, teaching practice. This covers a broad range of interventions that are mathematical in focus, from relatively 'small-scale' strategies, such as the use of representations, to 'large-scale' programmes that are intended to cover a large part of the curriculum offer in mathematics for a term or more. The critical characteristic is that the intervention is sufficiently well-described and could be implemented in KS3 and/or KS4 mathematics classrooms by schools and/or teachers in England (perhaps with some modification and in some cases with substantial costs). Thus, we exclude approaches that are not stable in definition (for example, studies where the intervention changes and develops over time, such as design research projects) and approaches that are not clearly and unambiguously distinct from usual practice. We also exclude studies where the focus is on understanding how pupils learn rather than on examining the efficacy of teaching approaches that enable pupils to learn. The terms interventions, approaches and strategies are used interchangeably. We define effectiveness in terms of mathematical attainment and, to be included in our review, studies of interventions are required to have a mathematical attainment outcome (although, in some cases, this may not be the outcome identified as primary by the study authors). Hence, we exclude studies where the focus is only on attitudinal, dispositional or behavioural changes (even where these are largely mathematical in focus) and studies where only measures of teaching practice, behaviours or competence are collected. In our previous review of KS2 and KS3 Mathematics (Hodgen et al., 2018), we identified a number of commonalities across approaches. Hence, we adopted a modular structure to the review and grouped the identified interventions, approaches and strategies into 24 'modules'. These modules were organised around the following broad themes:

- Pedagogic approaches;
- Resources and tools;
- Mathematical Topics;
- Wider school-level strategies;
- Attitudes and dispositions;
- Transition; and
- Teacher knowledge and professional development.

We will use this structure as a starting point for developing the topics for the present review. We will review our topics by March 2024 to ensure that any questions and challenges raised by the Secondary Mathematics Practice Review (Boylan, Forthcoming) are addressed (where appropriate) in our Evidence Review. In addition, we will conduct an

analysis of relevant meta-analyses identified using a systematic search strategy and will review the topics during the study screening process.

In commissioning this systematic review, the EEF have identified one additional topic of interest: the effectiveness of non-specialist teachers teaching mathematics. Given the recent increased interest in teaching approaches informed by cognitive science amongst mathematics education researchers, we expect to find more evidence here than for our previous reviews and, thus, to examine the evidence for such interventions and approaches as a separate topic. Our focus will be on high quality evidence (see Logic of the review) relevant to KS3 and KS4 mathematics teaching in England. We note that Malouf and Taymans's (2016) analysis of the What Works Clearinghouse (WWC) database indicated that, up to 2014, in mathematics, robust studies were relatively rare, and most teaching approaches had only a limited evidence base at best. Indeed, of the three recommendations in the What Works Clearinghouse (WWC, 2019) *Teaching Strategies for Improving Algebra Knowledge in Middle and High School Students* guidance, only one is supported by moderate evidence; the remaining two are supported only by minimal evidence. We noted a similar problem in our previous reviews (Hodgen et al., 2018; Hodgen et al., 2020b). Hence, it is likely that, as for other recent EEF guidance (and the various WWC practice guidance reports), some approaches of interest may still only be supported by limited robust experimental evidence. Here we will draw on a wider range of quantitative studies and, if necessary, consult an authoritative set of a priori agreed high-quality narrative syntheses (e.g., the US National Council of Teachers of Mathematics' Research Compendium: Cai, 2017). In some cases (for example, approaches to support non-specialist teachers of mathematics), it is likely that our findings will take the form of an evidence gap map (e.g., White et al., 2020) and focus on identifying plausible approaches for further investigation.

#### **[RQ2]. What are the key features of successful approaches for teaching mathematics in Key Stages 3 and 4?**

As described above, our analysis will consider the efficacy of both pedagogic and implementation components of effective interventions as follows:

*Pedagogic components:* We anticipate that the different approaches and interventions will incorporate a variety of pedagogic features or components (such as the use of representations, or structured practice) and that these features or components will be common to several interventions (both within and between topics). Given Seidel & Shavelson's (2007) finding that domain-specific components of teaching have a very much larger effect than generic components, our focus will principally be on mathematic-specific components. We will identify potential components in two principal ways. First, using the theory of mathematical learning outlined in our previous review of mathematics teaching at Key Stages 2 and 3 (Hodgen et al., 2018, 'Overview of the development of mathematical competency', pp. 16-26), we will identify potential features that are likely to improve mathematical learning and, hence, raise attainment. In doing this, we will examine pedagogic components directed at improving one or more of the five mathematical proficiencies identified by Kilpatrick et al. (2001): conceptual understanding, procedural fluency, strategic competence, adaptive reasoning and productive disposition. Second, we will identify further features through analysis and coding of the relevant meta-analyses identified through the systematic review process described below. This will be supplemented by an analysis of the authoritative set of high-quality narrative reviews agreed upon with the Advisory Group (see Advisory group, p.7, and Exemplary reviews, p.30). We will then code the various interventions in our dataset for these pedagogic components. During the coding process, we anticipate that some additional components may be identified. We will compare the relative effect of different individual components through meta-regression and, additionally, use QCA (Qualitative Comparative Analysis) (Thomas, O'Mara-Eves & Brunton, 2014) to identify clusters of components that appear to be associated with effective interventions.

*Implementation components:* In addition to identifying the likely effectiveness of the pedagogical features of different approaches, interventions and strategies for the teaching of mathematics, it is important to describe *how* they can be implemented by teachers in classrooms with pupils and how school leaders can support this implementation. The ways in which an 'effective' approach is understood and implemented will have a considerable impact on the actual effectiveness in practice. To do this, we will adopt a similar strategy to that outlined above for the pedagogical features and characteristics. Yet, this presents a significant challenge, because, in many experimental studies, the intervention is conducted in 'ideal' conditions and the process of implementation is inadequately described. It is likely that, as with Sims et al.'s (2021b) recent review of teacher professional development, we will identify a sub-set of studies with implementation and process evaluations (IPEs) of reasonably high quality (Maxwell et al., 2021) and we will identify implementation characteristics (such as the provision of professional development for teachers). It is likely that many of these studies will be of trials funded by the EEF or associated with the Institute of Educational Sciences (IES) in the US.



***[RQ3]. Do mathematics approaches have differential effects on outcomes for socioeconomically disadvantaged pupils (for example, those eligible for free school meals)? If so, what are the key features of successful approaches?***

Where possible, we will conduct sub-group analyses focusing on the effects (and differential effects) of approaches for socio-economically disadvantaged pupils (see, e.g., Steenbergen-Hu & Cooper, 2013; Steenbergen-Hu et al., 2016). Socio-economic status (SES) is measured in various ways. In England, SES is often captured at pupil-level through free school meals status (FSM), as in EEF trials, or through the Income Deprivation Affecting Children Index (IDACI), or a combination of the two. In other countries, SES may be measured in different ways, through similar measures of free lunch eligibility or parental income / occupation. Often, in the US, deprivation is measured at school-level and/or interventions are targeted at schools with deprived intakes. We will conduct separate sub-group analyses of studies using pupil-level SES measures and, if judged possible, studies with only school-level measures. In some systems, various indicators of disadvantage, such as gender, low attainment and SES, are conflated (e.g., Fuchs et al.'s, 2021, use of the term 'struggling learners'). Our study will treat SES as distinct to gender, low attainment or SEND status. In addition to SES, we will also examine any differential effects for other groups of pupils, including gender and low prior attainment. Gender is a particular issue, given the relatively low participation of girls in post-16 mathematics and beyond.

***[RQ4]. What is the evidence on the effectiveness of approaches that support the transition between Year 6 and Year 7 and between Year 11 and Year 12?***

We will examine transitions as one of our topics (focusing on the two key transitions in English secondary education: Year 6, primary, to Year 7, secondary, and Year 11, secondary, to Year 12, post-16). We note that we expect the research on the Year 11 to Year 12 transition to be more extensive than that for Year 6 to Year 7. Nevertheless, we anticipate that for both transitions, the relevant literature base is likely to be limited in scope and methodological rigour. We do not expect to find strong experimental studies. Hence, as in our previous review (Hodgen et al., 2018), we expect to draw on the high-quality narrative syntheses (see 'Exemplary reviews') in addressing this research question and that our findings will very likely indicate gaps in the evidence together with plausible approaches for further investigation.

***[RQ5]. What is the evidence on the effectiveness of non-specialist teachers teaching mathematics, and on support for non-specialist teachers?***

Non-specialist mathematics teachers may refer to qualified teachers of other secondary subjects teaching mathematics, to qualified primary teachers teaching mathematics in secondary schools, or to qualified teachers teaching mathematics in secondary schools who have neither a degree in mathematics or other numerate subject nor a mathematics-specific teaching qualification. We will examine evidence about the effectiveness of all three groups together with ways of supporting teaching by non-specialists (including but not limited to subject enhancement courses in mathematics). We note that there is likely to be a lack of experimental evidence, rigorous or otherwise. Hence, we propose to draw on the international 'grey' literature and wider narrative syntheses of research on non-specialist teachers (e.g., Goos et al., 2019; Hobbs & Torner, 2019) and to link this to the evidence on teacher knowledge and professional development. It is likely that much of this evidence relates to small schools or isolated rural schools and, thus, may not be wholly relevant to the problem of mathematics teacher shortages in England. Given the level of evidence that we anticipate, it is likely that we will draw on the high-quality narrative syntheses (see 'Exemplary reviews') in addressing this research question. As a result, our findings will take the form of an evidence gap map (e.g., White et al., 2020) and focus on identifying plausible approaches for further investigation.

## Advisory group

Across the review, we will draw on the advice of an Advisory Group at specific points. This advisory group will be made up of academics in mathematics education, those with expertise in systematic reviews, and practitioners. In addition to working with the Advisory Group, we will also collaborate with the Practice Review Team at Sheffield Hallam University, allowing us to respond to the developing concerns of practitioners in secondary mathematics education in England.

One role of our Advisory Group will be to provide guidance on developing an a priori set of authoritative narrative reviews relevant to secondary mathematics education in England. As discussed above (under RQ1), these reviews will be consulted where no, or very minimal, quantitative evidence exists in relation to a particular approach.

## Methodology

### Logic of the review

This review will take a sequential synthesis design (Noyes et al., 2019). To collate the most appropriate evidence to address each Research Question and to ensure we fully attend to the “different approaches” (described for the purpose of this review as “Topics” e.g., ‘*The use of representations*’) underpinning RQ1 (*What is the evidence on the effectiveness of different approaches for teaching mathematics in Key Stages 3 and 4?*), we will initially conduct ‘lumped’ (ibid, 2019) single sensitive searches to address all potential subtopics, followed by split searches (that is, searches focussed on a specific RQ or topic) of a broader literature base where RCT or high-quality QED evidence (see below) meeting our inclusion criteria for Phase 1 and pertaining to a particular RQ/topic is limited. This is outlined in the flowchart in Figure 1 1.

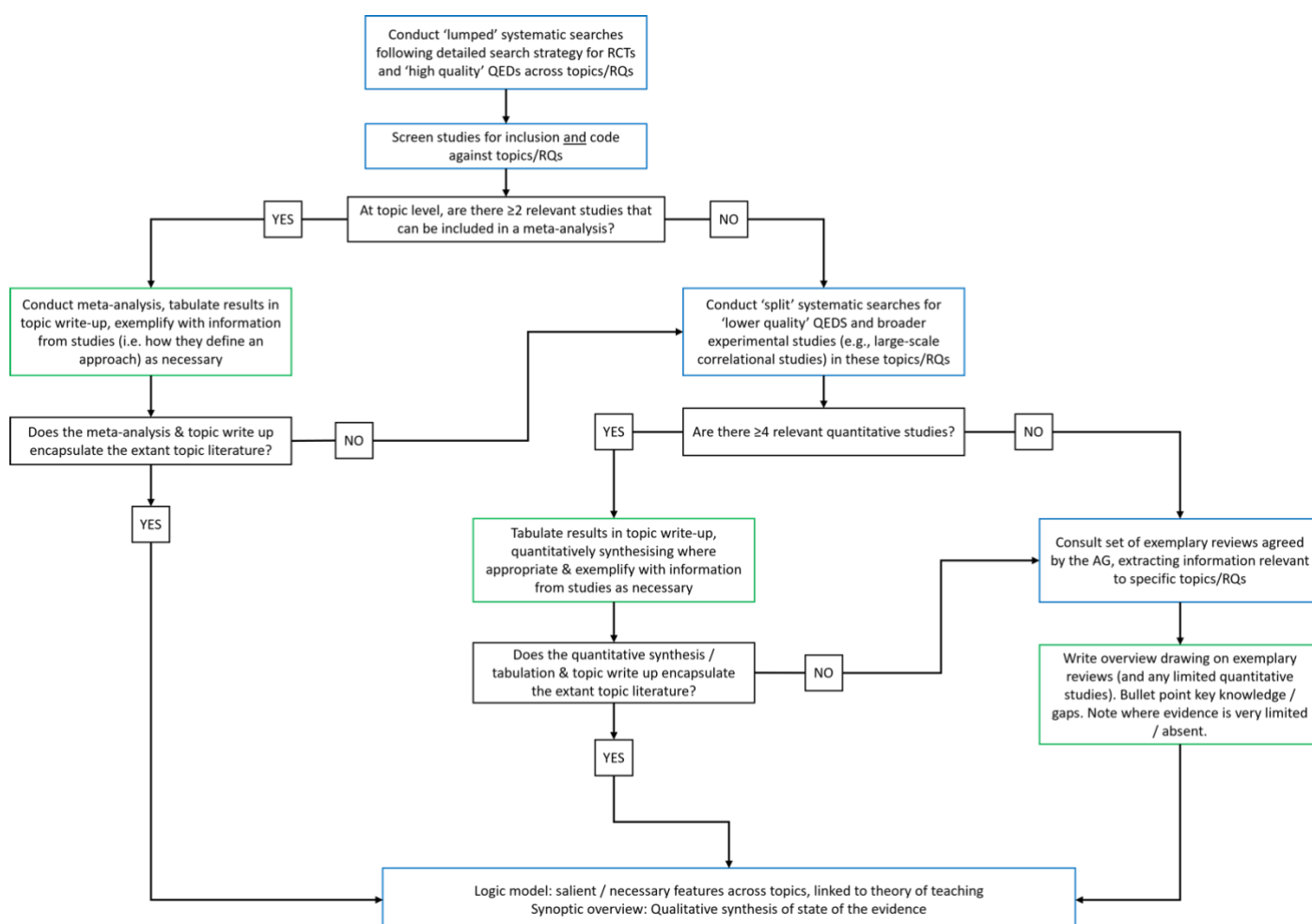


Figure 1: Flowchart of the logic of the literature review search process.

The literature search will therefore have two sequential phases:

1. A full systematic search (detailed below) for RCTs and QED studies of sufficient quality across all Research Questions and topics. Sufficient quality here refers only to the study design of the QED, with all QEDs with a concurrent ‘Business as Usual’ or active control group and a pre-test categorised as high-quality.
2. A systematic search for wider experimental and other quantitative studies (including QEDs not classified as ‘of sufficient quality’ such as those without a pre-test) of specific Research Questions and topics *where there are not (at least) two appropriate Phase 1 RCTs or high-quality QEDs that can be meaningfully combined*.

Where minimal or no evidence is found across both phases to support a RQ/topic (quantified as fewer than four relevant quantitative studies), we will consult a set of exemplary reviews (agreed a priori with the Advisory Group) to provide a

narrative commentary as to the current state of knowledge. This search approach will highlight where no or minimal evidence exists in relation to each RQ/topic, supporting discussion of future research directions and potential avenues of promise.

Outcomes from the sequential synthesis design will allow us to produce a consistent overview of the current evidence and understanding in relation to each topic, tabulating the quantitative evidence where appropriate. In addition to this – addressing RQ2: *What are the key features of successful approaches for teaching mathematics in Key Stages 3 and 4?* – we will identify the necessary and constituent parts of successfully implemented interventions. Further, we will bring the review together with a synoptic overview, taking a broad view of the current state of play, both in terms of what is known and the ways in which it is known, allowing us to call attention to future research directions and potential avenues of promise.

## Topic identification

As noted above, we conceptualise the “different approaches” underpinning RQ1 and the core themes of RQ4 (transitions) and RQ5 (non-specialist teachers) as “Topics” and use this terminology throughout. The terminology and approach of using topics builds on our use of “modules” in previous reviews (Hodgen et al., 2018; Hodgen et al., 2020b). We use topics to demarcate these from the Modules of the previous reviews which are built on here, showing the clear development of work across the reviews.

The list of topics for the present review will be developed in a four-part iterative process.

1. We will start with the list of Modules developed within Hodgen et al. (2018). This evidence-informed structured list of Modules is an appropriate starting point as the previous review covered Key Stage 3 and hence there is a cross-over in the age-phase and teaching approaches with the present review. We will add topics to this list to capture further aspects of the present review, e.g., ‘Non-specialist teachers’ in response to RQ5.
2. We will take guidance from the EEF Practice Review Team who are exploring the areas (which can be conceptualised as topics) of interest and importance to practising teachers.
3. Within our main systematic search (see Search strategy for identification of studies > *Searching for grey literature*) we will produce a full set of meta-analyses of interventions in secondary mathematics teaching published since 1968. This list will be coded in Excel against our developing topic list (from phase 1 and phase 2 above combined) with new topics added to reflect foci not in the current list.
4. During screening of studies within the main search process, studies marked for inclusion will be coded against the developing topic list (from part 1, part 2, and part 3 above combined) with new topics added to reflect foci not in the current list.

Studies will be coded to a primary and to a secondary topic. It is anticipated that this coding will, in part, support the development of our response to RQ2 as we will be able to assess which topics repeatedly co-occur.

This process will result in a full and expanded set of topics. We will review and discuss as a team where topics could be expanded or combined or where new ‘parent level’ topics could be added. This will enable us to compile our topic coding tool (See Appendix 2C) which will be used within the data extraction process in EPPI Reviewer (Thomas et al., 2023), enabling us to pull out studies for analysis related to specific topics. As with all coding tools in EPPI Reviewer, a 5% duplicate list will be coded against the topics to check for, and respond to as necessary, inter-coder reliability. The topic format will further provide the structuring for the main report, with discussion of the evidence presented on a topic basis for RQ1 and the topics used to support discussion in addressing other RQs.

## Information management

One database documenting studies identified in both phases of the literature search will be constructed in Microsoft Excel. This approach draws on the expertise of the team, allows for transparency in the search process, supports the removal of duplicates and allows for the integration of datasets constructed within previous studies (see Search strategy for identification of studies). Further, the field headers reflect those used in EPPI Reviewer, supporting transfer to the EPPI Reviewer platform at the data extraction stage (see Data extraction and management). Where publications contain

more than one potentially eligible study, each study will be recorded as a separate entry in the database, making each study the unit of interest under review.

The database will contain two sheets. Sheet 1 will be the main database. It will involve a full listing of studies identified across the search processes and all decisions (e.g., screening decisions) related to these. The final dataset for data extraction and synthesis will be pulled out from this database sheet. Sheet 2 will involve a full record of all searches and search hits and a record of the number of studies included, again allowing for transparency in the construction of the total corpus searched. The field headings and a description of each are given in Table 1 and Table 2. Further detail about the processes involved is given in the subsequent sections.

A	Record #	A unique identification number assigned to all studies, allowing for records to be cross-checked. This number will also be at the start of the filename for the associated sourced publication allowing for ease of identification and data management.
B	Date of entry	The date the record was added to this database.
C	Researcher conducting search	The member of the research team completing the initial entry following the search.
D	Search number	This will correspond with the search records in Sheet 2, allowing for full transparency as to where individual items have been sourced from.
E	Reference Type	Journal article / Conference Paper / Report / Dissertations and Theses / Unpublished.
F	Authors	All authors listed as on publication.
G	Publication Date	Published date as given (not received / online first date).
H	Title	Title of the publication.
I	Journal / Publisher	Name of journal, publisher, awarding institution (for dissertations and theses).
J	Volume	For journal articles only.
K	Issue	For journal articles only.
L	Pages	Where applicable.
M	ISSN	Where applicable.
N	DOI	Where applicable.
O	Abstract	Published abstract (or published summary, e.g., in reports).
P	URL	Where applicable.
Q	Full publication reference	Full reference in APA.
R	Multiple studies?	Indicates whether multiple studies (and how many) are being drawn from the same publication – a separate entry will be completed for each study.
S	Study name and #	Completed where multiple studies are drawn from the same publication.
T	Publication sourced?	<ul style="list-style-type: none"> <li>Y – full-text of the publication has been sourced and filed using the record # at the start of the filename for ease of data management.</li> <li>N – full-text unavailable (further details can be included in the 'notes' column, e.g. contact author to obtain copy)</li> </ul>
U	Name of screener	The member of the research team completing the screening processes.

V	Screen on title	<p>Recommendation following screening of the title only (see screening details below – this is an assessment of a limited number of inclusion criteria):</p> <ul style="list-style-type: none"> <li>• SCREEN ON ABSTRACT (Meets limited inclusion criteria for screening on title)</li> <li>• EXCLUDE (Does not meet all limited inclusion criteria for screening on title)</li> </ul> <p>This is an over-inclusive process designed to remove obvious studies which would not be suitable for inclusion, e.g., those specifying a population outside of our review.</p>
W	Screen on abstract	<p>Recommendation following reading of the abstract:</p> <ul style="list-style-type: none"> <li>• INCLUDE ON ABSTRACT (Meets all inclusion criteria)</li> <li>• SCREEN ON FULL-TEXT (Not possible to make a decision based on title and abstract – further details can be added to the ‘notes’ column where a specific feature, e.g., age of population, is queried)</li> <li>• EXCLUDE (Does not meet all inclusion criteria)</li> </ul>
X	Screen on full-text	<p>Recommendation following consultation of the paper:</p> <ul style="list-style-type: none"> <li>• INCLUDE ON FULL-TEXT (Meets all inclusion criteria)</li> <li>• EXCLUDE (Does not meet all inclusion criteria)</li> </ul>
Y-AC	Inter-coder checks	<p>Five hidden columns to allow for inter-coder reliability checks:</p> <ul style="list-style-type: none"> <li>• Name of second screener</li> <li>• Screen on title</li> <li>• Screen on abstract</li> <li>• Screen on full-text</li> <li>• Agreement – marked Y or N to calculate % agreement</li> </ul>
AD	Duplicate	<ul style="list-style-type: none"> <li>• 1 – no duplication (single copy of study)</li> <li>• 2 – duplication (record of study to be included)</li> <li>• 3 – duplication (record(s) of study to be excluded)</li> </ul>
AE	Study type	<ul style="list-style-type: none"> <li>• RCT</li> <li>• QED (with concurrent ‘Business-as-usual’ or active control <u>&amp;</u> pre-test)</li> <li>• QED (with concurrent ‘Business-as-usual’ or active control <u>without</u> pre-test)</li> <li>• Non-experimental design (with comparison group)</li> <li>• Natural experiment (with comparison group)</li> <li>• Before/after designs (no control group)</li> <li>• Correlational study</li> <li>• Cohort study</li> <li>• Other quantitative study</li> </ul> <p>This will enable appraisal of the literature base for each topic and allow for the study type dataset to be separated for analysis.</p>
AF	Key Stage 5 literature	Y – where the study population is equivalent to Key Stage 5 (post-16) (see further details in screening). Left blank if not marked as Key Stage 5.
AG	Notes	Highlighted based on actions required, e.g., contact study author for full-text.
AH	Topic coding	Columns to allow the study to be broadly attached to one or more RQ / topic

Table 1: Fields in main literature search database sheet (Sheet 1)

A	Search #	A unique identification number assigned to each search. This will be recorded against entries in Sheet 1 under column D to ensure transparency in the source of the studies.
B	Search Phase	<ul style="list-style-type: none"> <li>Phase 1 (RCTs and high-quality QED studies)</li> <li>Phase 2 (Wider experimental and other quantitative studies)</li> </ul>
C	Date of search	The date the search was conducted.
D	Researcher conducting search	The member of the research team completing the search.
E	Database, site, or source	The name(s) of the database(s), individual sites or source searched
F	Platform	The platform used hosting multiple databases
G	Search String	The search string used or full details of the search process
H	Hits	Total hits from specified search
I	Included on title	Number of publications included (on broad inclusion criteria – see search strategy) and copied to Sheet 1 for screening.
J	Notes	Highlighted based on actions required, e.g., contact study author for full-text.

Table 2: Fields in search record database sheet (Sheet 2)

## Inclusion and exclusion criteria for the review

A sequential synthesis design search strategy will enable us to collate publications (data) from which to directly extract evidence to address the RQs. The following inclusion and exclusion criteria, based on the PICOS model, will be applied to both phases of the search. Where appropriate (specifically for ‘Setting’, ‘Comparison’, ‘Outcomes’ and ‘Study design’) we specify the inclusion/exclusion criteria as applicable to each phase.

Criteria		Included	Excluded
POPULATION	Age	<p>&gt;50% of the sample, or an identifiable sub-sample, are learners or pupils in Key Stages 3 and/or 4, or school grade and pupil age equivalents as set out in the EEF Main Data Extraction Coding Guide (EEF, 2022a, p.37).</p> <p>The sample is predominantly Key Stage 2 learners or pupils, <u>and</u> the study covers transition into Key Stage 3</p> <p>The sample is predominantly Key Stage 5 learners or pupils, <u>and</u> the study covers transition from Key Stage 4</p>	<p>The majority of the sample, or all identifiable sub-samples, are: children under 5; primary school learners or pupils in Key Stage 1; primary school learners or pupils in Key Stage 2 (and not a study of transition); students in post-secondary education / post-16/KS5 (and not a study of transition)*; students in higher education; adults. <i>[Excluded as irrelevant to the focus of the study]</i></p> <p>Eligible sub-samples where an effect size is not reported or cannot be imputed.</p>

Additional needs	Low attaining pupils, including those broadly referred to as having dyscalculia, taught within mainstream educational settings.	Studies exclusively of pupils referred to generically as having 'Learning Difficulties/Disabilities' or identified as 'SEND' (or similar international term). Studies of a specified learning difficulty / difference (such as Williams Syndrome). <i>[Excluded as focus is on the difficulty/disability rather than on a mathematical approach applicable to all students].</i>
Setting / Context	<b>Phase 1 (RCTs &amp; 'high quality' QEDs):</b> The intervention or approach is undertaken in a mainstream school setting or represents part of the mainstream school 'offer' (such as homework). Interventions delivered by Teaching Assistants in mainstream settings where these supplement work done in class (the equivalent of <2 days of maths per week).	The intervention or approach is undertaken outside of the mainstream school setting e.g., studies conducted in laboratories, or studies of museum education, home-schooling, or special schools (including alternative provision). <i>[Excluded as findings / evidence cannot be implemented, or are complex to implement, by mainstream secondary school teachers or within mainstream secondary schools]</i> The intervention setting has a smaller granularity than a class, e.g., we will exclude studies with <15 in each group allocation. <i>[Excluded as a small sample size is one element of deciding that the study is about learning not teaching.]</i>
	<b>Phase 2 ('Lower quality' QEDs &amp; wider quantitative studies):</b> In addition to Phase 1 inclusion criteria, the intervention or approach may be undertaken in a laboratory setting and may be delivered by researchers.	The intervention or approach is undertaken outside of the mainstream school setting e.g., studies of museum education, home-schooling, or special schools (including alternative provision). <i>[Excluded as findings / evidence cannot be implemented by mainstream secondary school teachers]</i> The intervention setting has a smaller granularity than a class, e.g., we will exclude studies with <15 in each group allocation. <i>[Excluded as a small sample size is one element of deciding that the study is about learning not teaching.]</i>
Geographical location	The intervention is carried out in educational systems that at some point have taken part in PISA studies (see: <a href="https://www.oecd.org/pisa/aboutpisa/pisa-participants.htm">https://www.oecd.org/pisa/aboutpisa/pisa-participants.htm</a> and Appendix 1) or are associated with education systems that have taken part in PISA studies.	The intervention is carried out in educational systems that at no point have either taken part in PISA studies or are associated with systems that have taken part in PISA studies. <i>[Excluded to limit the searches to systems with high relevance to mathematics teaching in England]</i>

INTERVENTION	Intervention type	Clearly defined intervention (see our definition under RQ1). The intervention is well-described and could be operationalised by a teacher. The intervention or approach is mathematical; this may be as part of a wider cross-subject/school approach, but if so, there is a specific intervention within mathematics teaching going beyond the whole-school intervention. The intervention or approach is concerned with teaching (and teachers) or pedagogy (broadly, including e.g., assessment) or strategy.	The focus is solely on understanding how pupils learn mathematics (rather than on a teaching, or other, intervention aimed at improving pupils' learning specifically in mathematics). <i>[Excluded as such studies do not provide evidence of effective approaches]</i> Mathematics is an outcome, but the intervention does not address mathematics, e.g., a whole-school attendance intervention with an outcome measure of mathematical attainment. <i>[Excluded as this is not an intervention in mathematics]</i>
COMPARISON	Comparison or control conditions	<b>Phase 1 (RCTs &amp; 'high quality' QEDs):</b> Independent comparison group with control 'treatment' (i.e., 'business-as-usual' or active control) concurrent to the intervention.	No independent concurrent 'business-as-usual' or active comparison or control group. <i>[Excluded as not possible to measure gains as a result of a particular approach]</i>
		<b>Phase 2 ('Lower quality' QEDs &amp; wider quantitative studies):</b> May have a concurrent control group (but not necessarily) or be one-group studies.	
OUTCOMES	Outcome measures	<b>Phase 1 (RCTs &amp; 'high quality' QEDs):</b> Assessment of general – or an aspect of – mathematics attainment / achievement which reports quantitative continuous scores from testing of attainment / achievement / learning outcomes such as by standardised tests or other appropriate curriculum assessments or school examinations or appropriate cognitive measures (standardised or non-standardised)	No mathematics quantitative attainment outcomes measured – purely qualitative outcomes or attitudinal only outcomes. <i>[Excluded from main review as cannot be incorporated or aggregated into meta-analysis]</i>
		<b>Phase 2 ('Lower quality' QEDs &amp; wider quantitative studies):</b> Assessment of educational or cognitive attainment / achievement which reports quantitative continuous scores or dichotomous outcomes from testing (in any format, including researcher questioning) of attainment / achievement / learning outcomes such as by standardised tests, other appropriate curriculum assessments, researcher-constructed tools or school examinations or appropriate cognitive measures (standardised or non-standardised).	No mathematics quantitative attainment outcomes measured – purely qualitative outcomes or attitudinal only outcomes. <i>[Excluded from main review as cannot be incorporated into a quantitative synthesis]</i>



STUDY DESIGN	Study Design	<b>Phase 1 (RCTs &amp; ‘high quality’ QEDs):</b> Independent comparison group with control ‘treatment’ (i.e., ‘business-as-usual’ or active control) concurrent to the intervention <u>and</u> pre-test.	RCTs and QEDs without a concurrent ‘business-as-usual’ or active control group <u>and</u> a pre-test. <i>[Excluded as these would be categorised as Phase 2 studies]</i>
		<b>Phase 2 (‘Lower quality’ QEDs &amp; wider quantitative studies):</b> QED (with concurrent Business as Usual or active control but without a pre-test) Non-experimental design (with comparison group) Natural experiment (with comparison group) Before/after designs (one-group designs / no control group) Two (or more) treatment groups compared with no comparison or control Crossover designs without a mid-point test Discontinuity regression Delayed treatment (with/without pre-test) Correlational study Cohort study	Quantitative studies with inherent methodological flaws. <i>[Excluded as not possible to extract appropriate statistical outcomes for the intervention]</i>
	Publication Status / type	Published or otherwise publicly available literature in the form of journal papers, books or book chapters, working papers, reports, conference papers from published proceedings, theses and dissertations. Peer-reviewed or non-peer-reviewed studies.	Studies where the full-text is unobtainable (despite contacting the author / accessing the British Library). <i>[Excluded as inclusion is beyond the capacity of the research team and such studies are unlikely to be influential if not commonly available]</i>
	Language	English only	Not published in, or translated into, English. <i>[Excluded as inclusion is beyond the capacity of the research team]</i>
	Publication Date	Between 1968 and 01.02.2024.	Prior to 1968 and after 01.02.2024 <i>[Inglis &amp; Foster (2018, p.462) provide a justified rationale for “Accepting 1968 as a starting date for the modern research field” of mathematics education as the date when the journal ‘Educational Studies in Mathematics’ (ESM) was first published, noting the inception of the other preeminent mathematics education journal, ‘Journal for Research in Mathematics Education’ (JRME), two years later in 1970].</i>

Table 3: Inclusion and exclusion criteria for the review (*the text in orange font indicates our justifications for exclusion*)

\*While studies with a study population of post-16 / KS5 students are excluded from the study corpus (except for those focussing on transition from KS4 to post-16/KS5), any studies identified in our searches will be flagged in the database to support production of a separate dataset outside of this review for future analysis.

## Search strategy for identification of studies

As discussed in the Logic of the review, our review follows a sequential synthesis design with two phases. Each phase is discussed below. Results of searches and screening processes will be recorded as per Table 1 and Table 2, discussed in 'Information management'. The numbers of studies identified, then excluded / included across the search processes for each phase will be extracted from the Excel database into a PRISMA flowchart (Figure 2, p.23).

### Phase 1: Full systematic search for RCTs and high-quality QED studies across all Research Questions and Topics

We will begin the construction of our corpus for inclusion through conducting new systematic searches of the literature. These searches will be 'lumped' (Noyes et al., 2019) in nature, that is they search across all RQs and topics, rather than searching individually for literature pertaining to each RQ or topic separately.

#### *Search strings*

Search strings for the review are developed from the robust search strings used in Hodgen et al. (2020b), which in turn were developed from the search strings used in Hodgen et al. (2018) and Hodgen et al. (2020a). Search terms are updated to account for the age-ranges of concern in the present review, the foci of interest, and standard approaches in secondary mathematics.

Search terms are divided into five groups: general; subject, population; topic specific; and study design (Table 4). Search strings will cover all permutations across the five groups. For most database searches (see below) this will involve the use of Boolean Operators. As shown in the search terms table, wild cards (\*) will be used to account for different spellings and different suffixes, e.g., math, maths, mathematics, mathematical, mathematis, mathematising, mathematically, mathematician, mathematicians. The date-range will be set to search for content from 1968 onwards and the language to English, as per our inclusion criteria. We will allow for different language forms (randomized as well as randomised) in our searches.

Search levels will depend on the database being used. For most databases, searches will be conducted at the abstract level with a further search at the title level conducted where abstract returns are large, to reduce the hits and ensure later hits are not excluded. An example search using the ProQuest platform is given below:

```
abstract(approach OR education OR instruction OR intervention* OR learn* OR pedagogy OR programme OR strategy* OR teach*) AND  
abstract(arithmetic OR math* OR mathematic* OR numeracy OR calculus OR number OR algebra OR ratio OR proportion OR "rates of change" OR  
geometry OR measures OR probability OR statistics OR fluency OR reasoning OR "problem solving") AND abstract("11-16" OR "high school" OR  
"key stage 3" OR "key stage 4" OR "key stage four" OR "key stage three" OR KS3 OR KS4 OR "middle school" OR "secondary classroom*" OR  
"secondary education" OR "secondary level" OR "secondary school" OR "secondary teaching" OR "Year 7-11" OR "Grade 6-12" OR "Grade 6-8" OR  
"Grade 9-12") AND abstract(anxiety OR assessment OR attitude OR "blended learning" OR calculator* OR CGI OR "cognitive load" OR computer  
OR "concrete apparatus" OR "co-operative learning" OR "correspondence schools" OR CPD OR diagram* OR difficulties OR "digital tool*" OR "direct  
instruction" OR discuss* OR "executive function" OR "explicit instruction" OR "family engage*" OR feedback OR "generalist teach*" OR grouping OR  
group-work OR heuristic* OR homework OR imagery OR inquiry OR "integrative approaches" OR "isolated students" OR leadership OR manipulative  
OR mastery OR metacognition OR "misassignment of teachers" OR misconceptions OR "misplaced teachers" OR modelling OR motivation OR  
"parent* engage" OR PD OR "primary secondary transition" OR "professional development" OR real-life OR representation OR resource* OR "school  
leaving guidance" OR "school to work transition programs" OR "school visitation" OR "secondary postsecondary transition" OR self-instruction OR  
self-regulation OR setting OR "specialist teach*" OR structured OR "student adjustment" OR "student centred" OR "substitute teachers" OR task* OR  
"teacher background" OR "teacher competencies" OR "teacher distribution" OR "teacher placement" OR "teacher qualifications" OR "teacher shortage"  
OR technology OR textbook* OR track* OR traditional OR "transfer policy" OR transition OR "transition education" OR "transition programs" OR tutor*  
OR visualisation* OR whole-class OR working memory) AND abstract(RCT OR "randomised control trial" OR "randomised trial" OR "equivalence trial"  
OR "randomised experiment" OR QED OR "quasi-experiment*" OR "experimental design" OR "control group design" OR "control group trial" OR  
"comparison group design" OR "comparison group trial") AND stype.exact("Conference Papers & Proceedings" OR "Other Sources" OR "Government  
& Official Publications" OR "Reports" OR "Books" OR "Working Papers" OR "Scholarly Journals" OR "Dissertations & Theses") AND la.exact("English")  
AND pd(19680101-20240201)
```

Effective approaches to teaching mathematics in Key Stages 3 and 4  
Protocol for a systematic review and meta-analysis

General	Subject	Population	Topic Specific				Study design
approach education instruction intervention* learn* pedagogy programme strategy* teach*	arithmetic math* mathematic* numeracy number calculation calculus number algebra ratio proportion rates of change geometry measures probability statistics fluency reasoning proof problem solving model*	11-16 high school key stage 3 key stage 4 key stage four key stage three KS3 KS4 middle school secondary classroom* secondary education secondary level secondary school secondary teaching Year 7-11 Grade 6-12 Grade 6-8 Grade 9-12	anxiety assessment attitude blended learning calculator* CGI cognitive load computer concrete apparatus co-operative learning correspondence schools CPD diagram* difficulties digital tool* direct instruction discuss* executive function explicit instruction family engage*	feedback generalist teach* grouping group-work heuristic* homework imagery inquiry integrative approaches isolated students leadership manipulative mastery metacognition misassignment of teachers misconceptions misplaced teachers modelling motivation parent* engage	PD primary secondary transition professional development real-life representation resource* school leaving guidance school to work transition programs school visitation secondary postsecondary transition self-instruction self-regulation setting specialist teach* structured student adjustment student centred substitute teachers	task* teacher background teacher competencies teacher distribution teacher placement teacher qualifications teacher shortage technology textbook* track* traditional transfer policy transition transition education transition programs tutor* visualisation* whole-class working memory	RCT randomised control trial randomised trial equivalence trial randomised experiment QED quasi-experiment* experimental design control group design control group trial comparison group design comparison group trial

Table 4: Search terms for identifying literature within new systematic searches

### Databases and other search sources

Each search string will be run across the databases, repositories, and search engines listed in Table 5.

Database	Platform	Access provider
Australia & New Zealand Database	ProQuest	UCL
Ebook Central	ProQuest	UCL
Education Database (1988 – current)	ProQuest	UCL
Education Resources Information Center (ERIC) (1966 – current)	ProQuest	UCL
IBSS: International Bibliography of the Social Sciences	ProQuest	UCL
ProQuest Dissertations and Theses Global	ProQuest	UCL
Psychology Database	ProQuest	UCL
Social Science Database	ProQuest	UCL
Turkey Database	ProQuest	UCL
UK & Ireland Database	ProQuest	UCL
British Education Index	EBSCOhost	UCL
Child Development & Adolescent Studies	EBSCOhost	UCL
Education Abstracts	EBSCOhost	UCL
Education Index Retrospective: 1929-1983 (H.W. Wilson)	EBSCOhost	UCL
Educational Administration Abstracts	EBSCOhost	UCL
Humanities & Social Sciences Index Retrospective: 1907-1984 (H.W. Wilson)	EBSCOhost	UCL
OpenDissertations	EBSCOhost	UCL
Teacher Reference Center	EBSCOhost	UCL
Web of Science Core Collection	WOS	UCL
ProQuest Dissertations and Theses Citation Index	WOS	UCL
SciELO Citation Index (2002-present)	WOS	UCL
PsycARTICLES	Ovid	UCL
PsycEXTRA	Ovid	UCL
PsycINFO	Ovid	UCL
Australian Education Index	Stand-alone resource	UCL
BASE - Bielefeld Academic Search Engine	Stand-alone resource	UCL
Campbell Collaboration	Stand-alone resource	UCL
Digital Education Resource Archive	Stand-alone resource	UCL
EThOS (e-theses online service)	Stand-alone resource	UCL
JISC Journal Archives	Stand-alone resource	UCL
OpenGrey / DANS Data Station Social Sciences and Humanities (Archived – only updated to 2021)	Stand-alone resource	<a href="https://ssh.datastations.nl/dataverse/root">https://ssh.datastations.nl/dataverse/root</a>
SCOPUS	Stand-alone resource	UCL
CUREE—Centre for the use of evidence and research in education	<b>Research and Evidence Informed Leadership and Practice   Centre for the Use of Research &amp; Evidence in Education (CUREE)</b>	Open Access
Education Endowment Foundation: Completed projects, evaluation reports, database of studies	<b>Education Endowment Foundation   EEF</b>	Open Access
EIPPEE search portal – Evidence Informed Policy and Practice in Education in Europe	<b>Finding research: Search Portal (eippe.eu)</b>	Open Access
EPPI-Centre database of education research	<b>EPPI Centre Home (ioe.ac.uk)</b>	Open Access
JSTOR	<b>JSTOR Home</b>	Open Access
MRDC publications	<b>Publications   MDRC</b>	Open Access
What Works Clearinghouse – Institute of Education Sciences	<b>WWC   Find What Works! (ed.gov)</b>	Open Access
Google Scholar	<b>Google Scholar</b>	Open Access

Table 5: Databases, repositories, and search engines to be consulted.

Within Google Scholar we will apply key permutations of the search terms. Due to the over-inclusive nature of hits returned by this platform, we will limit scrutiny of the results to the first ten pages returned by each Google Scholar search.

### *Hand searches*

To support completeness in our study identification we will conduct hand searches of the below journals from May 2019 – February 2024. Our previous studies (Hodgen et al., 2018, Hodgen et al., 2020a) included hand searches up to April 2019, thus ensuring full date coverage. These journals are included in the hand search as they are the key journals where experimental studies in mathematics education are reported.

- Educational Research
- Educational Research Review
- Educational Researcher
- Journal for Research in Mathematics Education
- Journal of Educational Psychology
- Open Review of Educational Research
- Research in Mathematics Education
- Review of Education
- Review of Educational Research
- Review of Research in Education

### *Searching for grey literature*

To reduce the impact of publication bias we will include, as far as possible, grey literature in our data corpus, specifically theses and dissertations and unpublished reports. To identify these studies, we will consult the 'included studies' in all meta-analyses of interventions in mathematics education meeting our general search inclusion criteria and 'harvest' all studies which are grey literature and hence may not be picked up by our main searches.

A complete list of relevant meta-analyses will be developed in two ways:

1. From consulting the databases previously constructed for Hodgen et al. (2028) and Hodgen et al. (2020a) for meta-analyses which meet our inclusion criteria for the present review.
2. Through systematic searches.

The systematic searches for meta-analyses will take a similar approach to that outlined above for the main review searches. The search terms which will make up the search strings are given in Table 6. The same search databases, platforms and repositories (Table 5) will be used. An example of the search to be conducted on the Ebscohost platform would be:

AB ( approach OR education OR instruction OR intervention\* OR learn\* OR pedagogy OR programme OR strategy\* OR teach\* ) AND AB ( arithmetic OR math\* OR mathematic\* OR numeracy OR calculus OR number OR algebra OR ratio OR proportion OR "rates of change" OR geometry OR measures OR probability OR statistics OR fluency OR reasoning OR "problem solving" ) AND AB ( "11-16" OR "high school" OR "key stage 3" OR "key stage 4" OR "key stage four" OR "key stage three" OR KS3 OR KS4 OR "middle school" OR "secondary classroom\*" OR "secondary education" OR "secondary level" OR "secondary school" OR "secondary teaching" OR "Year 7-11" OR "Grade 6-12" OR "Grade 6-8" OR "Grade 9-12" ) AND AB ( meta OR "meta-analysis" OR "meta-analyses" OR "meta-analytic" OR "meta-syntheses" OR "analysis of analyses" OR "meta-evaluation" OR "meta-study" OR "meta research" OR metanalysis OR metastudy OR "research of researches" )

General	Subject	Population	Study design
approach education instruction intervention* learn* pedagogy programme strategy* teach*	arithmetic math* mathematic* numeracy number calculation calculus number algebra ratio proportion rates of change geometry measures probability statistics fluency reasoning proof problem solving model*	11-16 high school key stage 3 key stage 4 key stage four key stage three KS3 KS4 middle school secondary classroom* secondary education secondary level secondary school secondary teaching Year 7-11 Grade 6-12 Grade 6-8 Grade 9-12	meta meta-analysis meta-analyses meta-analytic meta-syntheses analysis of analyses meta-evaluation meta-study meta research metanalysis metastudy research of researches

Table 6: Search terms for identifying meta-analyses within new systematic searches

Sourced meta-analyses will be recorded in a separate Excel database following the same information management structure set out in Table 1 and Table 2. A third sheet within the same database will be used to record the harvested grey literature, linking this to the original meta-analysis from which it was sourced. Harvested grey literature will be added to the main literature database prior to screening and will hence undergo the same rigorous screening process applied to the full corpus.

### *Checking of Searches / Search Strategy*

To ensure our searches are comprehensive and identifying the full range of studies meeting our criteria and foci we will identify in advance a set of studies which should be identified for inclusion in the review through these searches. These will be blind to the team member carrying out the searches. We will then check that our searches identify each of these studies. Should any not be identified, the search strings will be expanded until the selected studies are identified, and all searches will then be re-run with the new expanded search strings.

### **Screening of Phase 1 literature**

Following Phase 1, all studies for screening will be contained in the main database (see Information management). We expect this to contain circa 2500 studies. Screening will involve the following phases and checks.

#### *Initial duplication removal*

The Excel database of all studies identified in Phase 1 (including harvested grey literature) will be sorted in turn by title, author, and full reference to exclude obvious duplicates. Duplicates will not be removed but will be marked using the duplicate coding system in Table 1. This supports transparency and enables the completion of the PRISMA flowchart for this search phase (Figure 2, p.23).

#### *Initial assessment (title only screening)*

Following duplicate removal, each study will be assessed at the title only level against a limited set of exclusion criteria drawn from the main inclusion and exclusion criteria, enabling the early removal of studies we would obviously not include.

Criteria		Exclusion Criteria
POPULATION	Age	The title states that the full focus of the study is: <ul style="list-style-type: none"> <li>• children under 5</li> <li>• primary school learners</li> <li>• Key Stage 1 higher education</li> <li>• adults</li> </ul>
	Additional needs	The title states that the full focus of the study is: <ul style="list-style-type: none"> <li>• Learners with EBD</li> <li>• Any specified learning difficulty (such as Williams Syndrome)</li> </ul>
	Setting / Context	The title states that the full study takes place in non-mainstream school settings: <ul style="list-style-type: none"> <li>• laboratory studies</li> <li>• museum education</li> <li>• home-schooling</li> <li>• special schools (including alternative provision)</li> </ul>
	Geographical location	The title clearly states that the intervention was wholly conducted in an educational system that at no point has either taken part in PISA studies or is associated with systems that have taken part in PISA studies.
INTERVENTION	Intervention type	The title clearly states that the intervention was in a subject or area which is not mathematical.
STUDY DESIGN	Language	The study is not published in, or translated into, English.
	Publication Date	The study was published prior to 1968.

Table 7: Exclusion criteria for title only screening

Title screening decisions will be recorded in the Excel database and marked only as either SCREEN ON ABSTRACT or EXCLUDE. All studies included at the title stage will go forward for abstract screening; no studies will be included on the basis of title screening only. Title screening will take an over-inclusive approach with any uncertainties marked for SCREEN ON ABSTRACT. Only where it is clear from the title (or publication date) that a study does not meet our criteria will it be excluded at this stage, e.g., when assessing whether the intervention is mathematical, a study would only be excluded at the title screening stage if it explicitly stated that it was an intervention in a non-mathematical area, e.g., Del Favero, L., Boscolo, P., Vidotto, G., & Vicentini, M. (2007). Classroom discussion and individual problem-solving in the teaching of history: Do different instructional approaches affect interest in different ways?. *Learning and Instruction*, 17(6), 635-657.

### Screening on abstract and topic coding

Each study passing screening at the title level will then be screened on its abstract. Here the full inclusion and exclusion criteria set out in Table 3, p.16, will be applied. Items will be marked in the Excel database as:

- INCLUDE ON ABSTRACT (Meets all inclusion criteria)
- SCREEN ON FULL TEXT (Not possible to make a decision based on title and abstract)
- EXCLUDE (Does not meet all inclusion criteria)

Assessments will be over-inclusive, that is, if the reviewer is unsure, the study will be marked for screening on full-text and carried through to the next phase of screening.

All studies marked as 'include on abstract' at the abstract screening stage will additionally be initially coded against the topic coding columns, again within the Excel spreadsheet. That is, if a study abstract indicates that the intervention involves worked examples in algebra, both 'worked examples' and 'algebra' will be checked (X) in the database. Further, the study type (RCT, QED (and type, etc.)) will be checked. Where a study is excluded at the abstract screening stage in this Phase as it does not meet our strict RCT / 'high quality' QED requirement, but would meet our inclusion criteria in Phase 2, the study design will also be entered here as this will enable us to revisit such studies in Phase 2 if necessary.

### *Obtaining full texts*

At this stage, full texts of studies marked for 'include on abstract' or 'screen on full text' will be sourced. Where the full text is unavailable, this will be noted in the database and the number of unavailable studies included in the PRISMA flowchart. These studies will also be coded as "exclude" for screen on abstract.

### *Screening on full-text and topic coding*

Each study marked as 'screen on full text' at the abstract screening level will then be screened through reading the necessary parts of the full text, e.g., description of the intervention, sample and/or methods. Here the full inclusion and exclusion criteria set out in Table 3, p.16, will be applied. Items will be marked in the Excel database as:

- INCLUDE ON FULL TEXT (Meets all inclusion criteria)
- EXCLUDE (Does not meet all inclusion criteria)

All studies marked as 'include on full text' at this screening stage will additionally be initially coded against the topic coding columns, again within the Excel spreadsheet as detailed above. This will result in all studies to be included in the review being coded against the topics. Further, the study type will be checked.

Where a study is excluded at the full-text screening stage as it does not meet our strict RCT / 'high quality' QED requirement, but would meet our inclusion criteria in Phase 2, the study design will also be entered – these items will be revisited in Phase 2.

Following this screening stage, the database will be cleaned, and a further duplication check conducted.

### *Inter-coder checks*

A random 5% sample of the post first duplication removal items (i.e., prior to screening on title) will be blind second screened by another member of the review team, following the same stages and processes outlined above. An inter-coder agreement score will be calculated, disagreements discussed, and any necessary adjustments which need to be made to the process will be instigated and subsequent processes re-run as required.

### *Forward citation searches of full-text includes*

Following screening we will conduct a process of forward-citation checking for all studies marked to be included. The purpose of the forward citation search is to identify any recent large and robust studies which may have been previously missed.

The forward citation search will involve searching for citations of studies published in high quality journals and included in our Phase 1 screening. The list of high-quality journals is developed from three sources:

- Journals rated as A\*, A or B in Toerner, G. and Arzarello, F. (2012). Grading Mathematics Education Research Journals, EMS Newsletter December 2012, 52-54. (Only those publishing in English)
- Journals publishing 10 or more articles submitted for REF 21 unit 23 and relevant to the foci of our review (available via Inglis et al. data sets at: [British Education Research and its Quality \(lboro.ac.uk\)](http://www.lboro.ac.uk/research/efee/research/efee-research/))
- Journals publishing robust articles included from all other searches in Phase 1 dataset

We will limit the studies we conduct forward citation searching on to those published from 2014 onwards, that is the date of the first EEF report and the date of the implementation of the current National Curriculum in England.



Those studies identified through the process above to be included in the forward citation search will be run through CitationChaser. Results will be screened to exclude:

- Anything published prior to 2014
- Non-journal publications
- Publications not in English
- Studies not published in journals in our high-quality journal list

The remaining studies will be tabulated as for all previous searches. Items will be screened on title as before. We will also check for duplicates against those items (both included and excluded) in the current database. Further, studies published after February 2024 will be excluded. All studies remaining will be included in the database and will then be subjected to the same screening process as with the main dataset.

Following the forward citation process, the PRISMA flowchart will be updated.

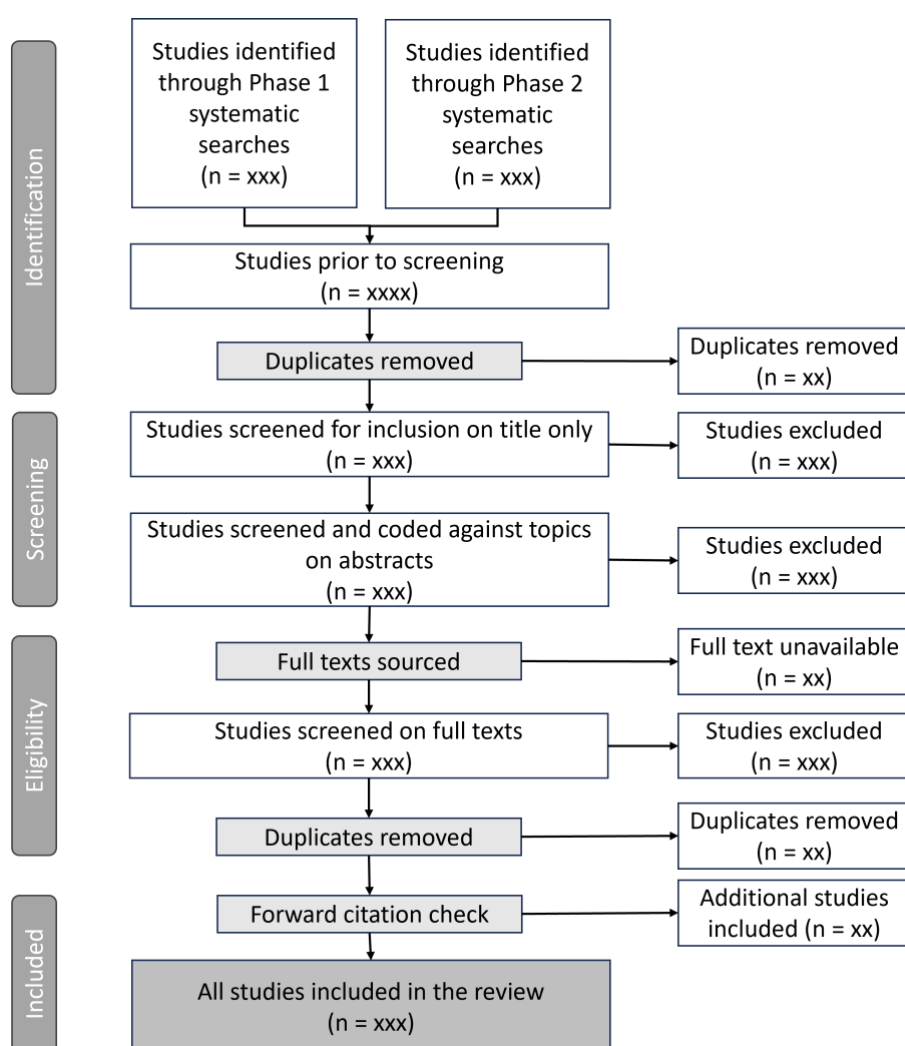


Figure 2: PRISMA flowchart of identified, excluded, eligible and included studies

### Phase 2: Systematic search for 'lower quality' QEDs and broader quantitative studies across selected Research Questions and Topics

Taking a sequential synthesis design enables us to conduct further highly specified searches only in particular areas (topics / RQs). Following Phase 1, we will have – through the topic coding applied during the screening processes, a

clear indication of the topics where RCT or 'sufficient quality' QEDs studies/evidence exist and the areas where this is minimal (defined as fewer than three studies) or non-existent.

At Phase 2 of our search design, we will only search for literature related to the topics identified as having no or minimal evidence. The search process will be almost identical to Phase 1, but with two important differences:

- Rather than taking the form of a 'lumped' search (i.e., searching across topics) we will conduct individual split searches with key words related to a specific topic or RQ. Where appropriate, subject heading searches (utilising the consistent vocabulary within particular databases) will additionally be conducted.
- Search terms related to the study design will allow us to identify 'lower quality' QEDs and wider quantitative studies as required. These terms are those used in the vocabulary lists of key databases to pick up the study designs in our inclusion criteria table (p.15). Search terms here will be split into two groups, with the search first run with group (i) terms, and then re-run with group (ii) terms only in cases where group 1 search terms return insufficient literature:
  - i. non-experimental design; natural experiment; difference-in-difference; regression-discontinuity; interrupted time series; delayed-treatment; matched design; comparison-group study; cross-over
  - ii. pre-post test design; before/after design; one-group design; correlational study; cohort study; relationship; test

An example search-string to be run on the Proquest platform within this phase would be:

abstract("misassignment of teachers" OR "misplaced teachers" OR "teacher distribution" OR "teacher competencies" OR "teacher placement" OR "teacher qualifications" OR "teacher shortage" OR "substitute teachers" OR "teacher background" OR "correspondence schools" OR "isolated students" OR "specialist teach\*") AND abstract("non-experimental design" OR "natural experiment" OR "difference-in-difference" OR regression-discontinuity OR "interrupted time series" OR "matched design" OR "cross-over") AND stype.exact("Conference Papers & Proceedings" OR "Other Sources" OR "Government & Official Publications" OR "Reports" OR "Books" OR "Working Papers" OR "Scholarly Journals" OR "Dissertations & Theses") AND la.exact("English") AND pd(19680101-20240201)

Search and screening results would be entered into the same database as before. Information on the numbers of studies sourced, screened, checked for eligibility, and included in the review will be entered into the PRISMA flowchart.

### Exemplary Reviews

As our 'logic of the review' flowchart (Figure 1, p.8) shows, we expect there to be some topics / RQs where searching at Phase 1 and at Phase 2 fail to produce satisfactory quantitative evidence (conceptualised as fewer than four relevant quantitative studies) to enable us to produce a meaningful synthesis (meta-analysis of broader exemplification of the data). In these cases, we will consult a prior agreed set of exemplary narrative reviews in mathematics education to enable us to provide commentary on the topic / RQ, supporting discussion of future research directions and potential avenues of promise.

The agreed set of exemplary narrative reviews will be developed in agreement with the Advisory Group, drawing on our own knowledge of the literature, advice from the Advisory Group, and non-systematic searches for highly cited narrative reviews including searches of Government websites such as the Ministry of Education and related websites in Australia, Ireland, New Zealand, Northern Ireland, Scotland, Wales, and others.

## Data extraction and management

### Importing studies and pre-coding checks in EPPI Reviewer

Following all screening and eligibility checks, we will produce a final database of all studies (located from both Phase 1 and Phase 2) to be included in the study. References in this database will be converted to RIS format and uploaded to EPPI Reviewer. Using a combination of Zotero functionality and manual uploading, full texts of each study will be attached to each reference.

References will be randomly allocated, using the EPPI Reviewer functionality, to the coding team. Each imported study will be checked by the coder as per the pre-coding checks outlined in the EEF Main Data Extraction Tool Guide (EEF, 2022a). Missing publication information (Author, title, journal, year, abstract, etc.) will be added.

## Data extraction tools

Data will be extracted and recorded within EPPI Reviewer from the selected papers, using a coding framework based on three tools:

1. The EEF main data extraction tool (version 2) (EEF, 2022a)
2. The EEF effect size data extraction tool (version 2) (EEF, 2022b)
3. Topic and intervention evaluation coding tool

Our three data extraction tools detailed above are provided in Appendix 2. All studies identified under Phase 1 of the literature search (that is, RCTs and 'high quality' QEDs) will be fully coded against each tool. Literature identified under Phase 2 ('lower quality' QEDs and other quantitative studies) will be coded where possible / sensible against the EEF coding tools and fully against Tool 3. Tool 3 will enable us to both extract data relevant to each topic and identify cross-cutting features of successful interventions, supporting us in addressing RQ2 (see also 'Responding to a lack of quantitative evidence: Exemplary reviews

There will be topic areas/RQs where quantitative evidence is scant or non-existent ( $\leq 4$  studies reporting quantitative outcomes). Where very limited or no quantitative evidence exists, we will use a set of exemplary or authoritative narrative reviews and practitioner guidance (e.g., Cai, 2017; Star, 2015), identified in advance in consultation with the Advisory Group, to succinctly outline what is known and where the gaps in knowledge exist in relation to that topic/RQ. It is likely that our findings will take the form of an evidence gap map (e.g., White et al., 2020) and focus on identifying plausible approaches for further investigation.

Key features of successful approaches'). There is some intentional duplication between Tools 1 & 2 and Tool 3 to ensure consistency in data capture across the corpus (both Phases 1 and 2), supporting us in consistently addressing topics with substantial evidence and those with more limited evidence.

Data extraction will be conducted by two trained Research Assistants. A random sample of at least 25 studies will be double-coded by other members of the research team. This applies to all extraction items (such as those relating to quality, 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 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.

## Assessment of risks of bias

In addition to the data extraction tools discussed above, we will also extract data and code to allow us to identify the potential for biased effect size estimates in individual included studies. This tool will be applied to all studies identified in Phase 1 and to applicable studies (that is those reporting an effect size estimate) identified in Phase 2.

The coding tool related to this appraisal is included in Appendix 2d. This tool was developed from the "assessment of risks of bias" coding included in Sims et al. (2021a) which is generally appropriate for the present review, with some amendments as justified below:

1. Is any outcome data missing for any reason by allocated group (attrition)? (Meets WWC Conservative Threshold, Meets WWC Liberal Threshold, Does not meet WWC Liberal Threshold, Extremely high [beyond WWC scale >58%], Not reported) [Justification: WWC thresholds take account of both overall and differential attrition in one scale and, hence, judged to be more appropriate]
2. Is compliance evaluated and reported? (Quantitative evidence reported, Only qualitative evidence reported, Not reported) [Justification: % attendance inappropriate metric for classroom interventions, very few studies report and compliance evidence]
3. Was the analysis pre-specified? (1; Not reported) [Justification: Assume not pre-specified if not reported or cited in publication. Very few studies report pre-specified design & methods]

4. Was the experiment analysed at the same level as it was randomised? (0/1; Not reported)
5. What was the number of units randomised? (Sample size)

As we include QEDs (of both 'high' and 'lower' quality) our tool also codes QED studies (of medium or higher relevance) for baseline equivalence: Are members of the treatment group the same as members of the comparison group before the study began? (0/1; Not reported).

### Effect size calculation (for missing ES)

We will extract and record all effect sizes using the Effect Size Data Extraction tool (Appendix 2b & EEF, 2022b). Outcomes will be based on continuous test scores and we will calculate standardised mean difference effect sizes as Cohen's *d*. Where effect sizes are not reported, or not reported in a format suitable for extraction into EPPI Reviewer, we will use the four calculators (in order – i.e., Campbell Calculator if applicable) recommended in the EEF effect size data extraction guide (EEF, 2022b) to calculate an effect size (as Cohen's *d*) or convert a reported effect size to Cohen's *d*:

- The Campbell Calculator [<https://www.campbellcollaboration.org/research-resources/effect-size-calculator.html>]
- Lenhard & Lenhard's Psychometrica [[http://www.psychometrica.de/effect\\_size.html](http://www.psychometrica.de/effect_size.html)]
- Lee Becker Effect Size Calculator [<https://lbecker.uccs.edu/>]
- The `escalc()` function in the metafor R package

In cases where an ES is provided but without a SE or CIs, a calculator based on WWC procedures will be used, as advised and supplied by the EEF Toolkit team at the University of Durham. This calculator was based on sample size of students and does not take account of clustering.

Where calculation is necessary, effect sizes will generally be calculated using means and standard deviations, confidence intervals or standard errors. Where these are not available, we will use, in order of preference, *t*-values or *F*-statistics, or *p*-values. Any results that can be converted to an effect size will be used. We will correct Cohen's *d* for bias in studies using Hedges' *g*. To provide a transparent record for effect sizes logged in EPPI Reviewer, any calculations undertaken on other websites or on a spreadsheet will be recorded, with a labelled screenshot of the calculation as detailed in the guide (EEF, 2022b, p.25).

We will use one effect size for the overall experimental group (or identifiable sub-group(s) of KS3/4 age). In cases for which multiple attainment outcomes are specified, we will by preference use the most general and most robust mathematics attainment outcome. Standardised measures are to be preferred over researcher-designed measures. Where multiple relevant attainment outcomes are reported and no primary outcome is specified, we will select an outcome at random.

We expect the number of studies where multiple mathematical attainment effect sizes are reported to be very few. In such cases, we will use the most general outcome of mathematical attainment and, if there is more than one candidate, we will choose the outcome randomly. As a sensitivity analysis, we will examine the effect of including multiple dependent measures using Pustejovsky & Tipton's method (2022).

### Unit of analysis issues

We will record the level of randomisation for randomised trials during the data extraction process using the standard EEF data extraction codes. Our expectation is that most studies will be randomised at a group level (e.g., classes or schools) rather than individual level. In order to account for these different levels of randomisation on the effect size estimates and variances, for studies where randomisation occurs at the group level (e.g., classes or schools) rather than the individual pupil-level, we will use the White and Thomas (2005) adjustment for clustering.

## Dealing with missing data

We will record missing, or unreported, outcome 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. We will conduct a sensitivity analysis for all calculated effect sizes using sub-group analysis.

## Data synthesis

Here we outline the methods to be used to synthesise our data which lead into our review reporting.

### Meta-analyses and sensitivity analysis

Our initial analysis will take place at a topic level, responding to RQ1 as well as addressing the themes of RQ4 (transitions) and RQ5 (non-specialist teachers). Where appropriate evidence exists ( $\geq 2$  effects sizes from RCTs or 'high quality' QEDs which can be sensibly synthesised) we will extract coded data from EPPI Reviewer to conduct meta-analyses within each topic/RQ. Meta-analyses will be conducted in the *metafor* package in R (Viechtbauer, 2010), using a random effects model to account for study variance and allowing us to include all relevant effect sizes in the analyses. We will use an inverse variance weighting approach to account for sample size (Deeks, Higgins & Altman, 2020). We will present the results in a forest plot together with the overall mean effect size (where appropriate). If there is statistically significant heterogeneity (see section 'Investigation of heterogeneity' below), then we will not report the overall mean effect size (i.e., the 'diamond' in the forest plot), but will report only the effect sizes and confidence intervals for each study and the heterogeneity statistics in the forest plot. Whether or not there is statistically significant heterogeneity, we will conduct analyses to explore possible sources of variation in the studies (see Investigation of heterogeneity below).

We will investigate the influence of potential outlying studies using the `influence()` and `leave1out()` functions in *metafor*. We will investigate potential publication bias by employing 'Trim and Fill', model selection analysis and the use of funnel plots.

We will conduct sensitivity analyses using the data coded in our Assessment of risk of bias coding tool (Appendix 2d) to check whether the results vary based on the categories of our appraisal of study design, namely:

- Attrition
- Non-compliance
- Form of comparison group
- Design (RCT or QED)
- Pre-specification
- Analysis level
- Units randomised
- Type of outcome test
- Baseline equivalence (for QEDs)

Pseudo code indicating the methods of analysis that we will use is included in Appendix 3.

### Sub-group analysis

Where the size and make-up of the topic dataset allows, we will conduct sub-group analyses to establish any differential impact of each approach / intervention on disadvantaged pupils, by gender, and by ethnicity (hence addressing RQ3).

### Investigation of heterogeneity (within and across topics)

We will calculate the following tests/statistics to determine the presence of statistical heterogeneity: Q-test,  $I^2$  and tau-squared. If the Q-test result is statistically significant ( $p < .05$ ), then this will be an indicator of significant heterogeneity (Lipsey & Wilson, 2001). If the  $I^2$  exceeds 75%, then this will be considered 'considerable heterogeneity' (Deeks, Higgins & Altman, 2020) and no overall mean effect size will be reported for the given analysis.

We anticipate that there will be some conceptual/characteristic heterogeneity that will be useful to explore and will support us in addressing RQ2 (*What are the key features of successful approaches for teaching mathematics in Key Stages 3 and 4?*). At a modular level, we will use sub-group meta-analysis to explore whether effect sizes vary based on characteristics of the implementation of the intervention, for example the inclusion of a specified programme of professional development. Such moderators will be extracted from our coding against Tool 3 (specifically, 2.5: inherent features in the implementation of the intervention) (Appendix 2c) rather than specified in advance, allowing us to identify key characteristics and their impacts.

Going beyond the modular level, and dealing with the complexity in the literature where interventions may overlap multiple topics, we will report an overall or unmoderated random effects mean for interventions in mathematics in the manner conducted by Gersten et al. (2009) in their meta-analysis of mathematics interventions for students with learning disabilities, allowing us subsequently to conduct a meta-regression to assess the general characteristics of interventions (that is, the characteristics of the implementation of the intervention discussed above) which appear to make them successful (RQ2). Further, we will use QCA (Qualitative Comparative Analysis) to investigate common clusters of features that appear to be associated with effective interventions (Thomas, O'Mara-Eves & Brunton, 2014).

### **Synthesising quantitative data for topics with limited evidence**

From our knowledge of the literature and previous studies (Hodgen et al., 2020b; Hodgen et al., 2018) we anticipate there being modular areas or research questions where an appropriate level of RCT or high-quality QED evidence ( $\geq 2$  appropriate studies) does not exist. In these areas, we will turn to our Phase 2 collated literature, that is 'lower-quality' QEDs and other quantitative studies.

Quantitative synthesis here will vary, based on the available evidence. Where appropriate (based on study design and reported statistical information) we will conduct a meta-analysis within these topics as outlined above. We do however anticipate this not being possible with the current evidence base for some topics/RQs. In such cases we will be led by the available evidence, synthesising outcomes if appropriate, providing the range of outcomes, or tabulating and providing a commentary on the outcomes individually and as a whole, including commentary on study design limitations such as sample size.

## **Reporting**

The reporting of our findings will have the following parts:

1. An overall synopsis of the impact of interventions applicable to secondary mathematics education in England, outlining the state of evidence
2. An account of the data and evidence for each topic, including sub-group (e.g., SES) analysis (RQ1; RQ3)
3. Analysis of the key features of successful interventions for teaching mathematics in Key Stages 3 and 4 (RQ2)
4. Analysis of the data and discussion of the evidence pertaining to transitions (RQ4)
5. Analysis of the data and discussion of the evidence pertaining to non-specialist teachers (RQ5)

Our synoptic overview will provide commentary drawing on evidence from across topics, in addition to presenting our overall effect size, to give a broad understanding of the state of the evidence as it currently exists in relation to successful approaches in mathematics teaching in mainstream secondary schools in England.

### **Modular (RQ1) and RQ4 / RQ5 reporting**

#### Narrative Reporting

For each topic, we will tabulate quantitative results of our main and sub-group analyses (as appropriate as discussed above). This will include use of previously developed expert-judgement approaches to assess the methodological quality of studies as well as the strength of evidence and directness (or relevance) of our findings for KS3/4 classrooms in England, based on the GRADE (Guyatt et al., 2008), Campbell Collaboration (2016), and AMSTAR (2021)

methodological quality tool approaches (see below). We will produce a narrative analysis for each topic, reporting on the evidence base to support sense-making and contextualisation of the quantitative outcomes. This will include a headline overview, definitions, commentary on findings (particularly homo- or heterogeneity of studies, the quality of evidence, relevance to English KS3 and KS4 mathematics education), issues relating to the implementation of the approach, and links to other topics.

### Quality of evidence

Within each topic, we will assess the quality of individual studies, with each study placed into one of six categories: excellent, very high, high, medium, low, or very low, based on the design, scale of the trial, level of attrition, the quality of the analysis, whether a standardised test was used, whether compliance was reported and whether there were any significant threats to validity. The procedure for assessing quality and assigning studies to a quality category are set out in Appendix 4.

At a topic level, we will assess the overall quality of the evidence within each topic. We will apply an adapted version of our previously developed expert-judgement approach (Hodgen et al. 2020b) to quality appraisal to assess the methodological quality of the evidence overall based on the GRADE (Andrews et al., 2013a, 2013b; Guyatt et al., 2008), Campbell Collaboration (2016), and AMSTAR (2021) methodological quality tool approaches. Table 8 details how the review team will make judgments about the quality of the body of evidence for each topic, and the extent to which the findings are supported by a robust body of evidence. The term ‘quality’ is used in preference to ‘strength’ to avoid confusion with the size of effects. For the first two topics, the full core team (three members) will make independent judgments, which will then be compared, aggregated, and moderated. Disagreements will be discussed and resolved as a team. Once an understanding is reached, judgements going forward for the remaining topics will be made by two core team members and moderated in the same manner.

Aspects of quality of evidence	Comments	Grade [0, minimal, to 3, strong]	Notes
A: The number of original studies	Summary of studies: number & scale		<ul style="list-style-type: none"> <li>Thresholds: 20 studies, strong [3]; 5 or less, low [1]; and none as minimal [0]</li> <li>For strong grade, at least 5 studies conducted at scale (&gt;250 pupils in the treatment group)</li> </ul>
B: The methodological quality of the original studies	Studies will be grouped into four categories: very high, high, medium, low		<ul style="list-style-type: none"> <li>Thresholds: For strong [3], at least 3 studies are of very high or excellent methodological quality; For medium [2], at minimum of 3 studies are of at least high methodological quality; For low [1], a minimum of 1 study is of at least high methodological quality; otherwise as minimal [0].</li> </ul>
C1: Consistency of results across the studies: effects	Consistency of effects (meta-analysis & homogeneity)		<ul style="list-style-type: none"> <li>Is there any evidence of heterogeneity? (If there is statistically significant evidence of heterogeneity, we will not report an overall aggregated ES.)</li> </ul>
C2: Consistency of results across the studies: intervention descriptions	Consistency of intervention descriptions		<ul style="list-style-type: none"> <li>Is the intervention sufficiently similar (and coherently described) across the studies? / Is any heterogeneity identified in C1 sufficiently well explained?</li> </ul>

D: Any reporting bias	Meta-analysis publication bias		<ul style="list-style-type: none"> <li>Is there any indication (or evidence) of publication bias?</li> </ul>
E: Evidence from existing systematic reviews and Best Evidence Syntheses	Alignment		<ul style="list-style-type: none"> <li>Do the narrative reviews support the findings of the original studies?</li> <li>If not, are there good reasons for the differences?</li> </ul>
Overall judgment of the strength of evidence	Justification and any caveats		Make overall judgment based on above criteria, then moderate across the team.

Table 8: Quality of evidence judgement tool

### Relevance of studies to mathematics teaching in England's secondary schools

As with our appraisal of methodological quality, we will make a judgement about the relevance for teaching mathematics in Key Stage 3 and Key Stage 4 in England at both an individual study and at a topic level.

To process what is expected to be a significant corpus of studies, particularly in relation to certain topic areas, and to ensure the reporting is useful and has relevance to mathematics teaching in England in Key Stage 3 and Key Stage 4, we will assess the relevance of each study. Studies will be grouped into four categories: very high, high, medium, and low, based on the extent to which the educational context is relevant to mathematics classrooms in England and the scale on which the intervention was trialled (to ensure that the intervention was described in a way that could be operationalised by class teachers in England). Studies with a significant threat to validity due to their publication status (e.g., those published in predatory journals (see Beall's List), inappropriate journals or which are MA theses) will automatically be placed in the low relevance category.

To ensure consistent judgements about individual study relevance across the team, a flowchart (See Appendix 5) to ascertain the appropriate relevance category will be used and these categories recorded within the EPPI Reviewer coding. Topic reporting (see below) will focus on those studies deemed to be of Very High, High, and Medium relevance, with Low and Very Low Relevance studies used as a checkpoint.

At a topic level, we will assess the relevance of the studies within the topic as a whole, judging the extent to which the available evidence is relevant to teaching mathematics in Key Stage 3 and Key Stage 4 in England. Table 9 details how the review team will make judgments about relevance. As with our quality appraisal, for the first two topics, three members of the review team will make independent judgments, which will then be compared, aggregated, and moderated. Disagreements will be discussed as a team. Following this, two members of the team will reach a moderated agreement for the remaining topics. Relevance is not independent of the quality of the body of evidence, so the overall relevance grading cannot be more than one grade higher than the quality of evidence grading, although we anticipate that in most cases relevance grading is likely to be lower than quality grading.

Aspects of relevance	Grade [0, minimal, to 3, high]	Notes
A: Where and when the studies were carried out		<ul style="list-style-type: none"> <li>Were any studies carried out in England? For strong grade, at least 1 study of very high relevance (conducted in the UK).</li> <li>Were the studies carried out in educational systems or contexts judged to be similar to England (either similar overall or similar for the topic)?</li> <li>If mostly US, is this aspect of US mathematics education judged to be sufficiently similar to England to be relevant?</li> <li>If many of the studies are dated, is this a threat to relevance? (to be assessed on a modular basis –</li> </ul>



		e.g., dated technology studies may be more problematic)
B: The number of studies assessed as of high and medium relevance		<ul style="list-style-type: none"> <li>• Thresholds: For strong [3], at least 5 studies are of at least high relevance; For medium [2], at minimum of 3 studies are of at least high relevance; For low [1], a minimum of 1 study is of at least high relevance and a minimum of 5 studies are if medium relevance; otherwise as minimal [0].</li> </ul>
C: How the interventions were defined and operationalised		<ul style="list-style-type: none"> <li>• Are the interventions either available in England or sufficiently well-described to be adapted for teachers to implement in England?</li> <li>• Are there widely available examples of use in England (although the particular interventions may not have been subject to a robust experimental evaluation)?</li> </ul>
D: Any focus on particular topic areas		<ul style="list-style-type: none"> <li>• Are the studies skewed towards particular mathematical topics – both broad topics (number/calculation v shape/space/geometry v measures) and more specific (narrow) topics?</li> </ul>
E: Age of children /phase of education		<ul style="list-style-type: none"> <li>• Were the studies carried out across the age range?</li> <li>• Are there reasons why the intervention is more appropriate for either KS3 or KS4?</li> </ul>
F: Ease of implementation		<ul style="list-style-type: none"> <li>• Are there potential difficulties with implementation (e.g., cost, amount of training required, level of external support required)?</li> </ul>
Overall relevance judgment		Make overall judgment based on above criteria, then moderate across the team. Focus more attention on criteria A and B with C, D and E as caveats.

Table 9: Relevance of evidence judgement tool

Quality and relevance will be reported as tables for each topic.

### Responding to a lack of quantitative evidence: Exemplary reviews

There will be topic areas/RQs where quantitative evidence is scant or non-existent ( $\leq 4$  studies reporting quantitative outcomes). Where very limited or no quantitative evidence exists, we will use a set of exemplary or authoritative narrative reviews and practitioner guidance (e.g., Cai, 2017; Star, 2015), identified in advance in consultation with the Advisory Group, to succinctly outline what is known and where the gaps in knowledge exist in relation to that topic/RQ. It is likely that our findings will take the form of an evidence gap map (e.g., White et al., 2020) and focus on identifying plausible approaches for further investigation.

## Key features of successful approaches (RQ2)

We will address RQ2, identifying the key features of successful approaches for teaching mathematics at Key Stages three and four, by examining characteristics associated with both the pedagogic aspects of the approach (such as, the use of representations) and factors associated with implementation factors (e.g., Lendrum & Humphrey, 2012). We will do this in two systematic ways.

Firstly, results from our investigation of heterogeneity across topics will show whether effect sizes vary based on characteristics of the implementation of the intervention (with characteristics extracted from coding against 2.5 of Tool 3, Appendix 2C). This will allow us to identify whether particular characteristics of different interventions have contributed

to different effect sizes. Looking across topics, we will be able to see whether certain characteristics are consistently implicated in the heterogeneity of interventions and the size and direction of any such effect.

Secondly, through our coding in Tool 3 (Appendix 2C) we will extract those studies (across all interventions / topics) where we assess the intervention to be both (a) successful and (b) particularly well-described with inherent features/characteristics identified by the study authors. We will tabulate (constructing a 'truth table') and classify salient implementation and design characteristics across these successful interventions. Using aspects of the methodological approach of Qualitative Comparative Analysis (Thomas, O'Mara-Eves & Brunton, 2014) we will then "identify those configurations of participant, intervention and contextual characteristics that may be associated with a given outcome" (ibid, 2014, p.1). This will enable us to highlight what needs to occur – and potentially when – within an intervention for its outcomes to be more likely to be successful.

Following the above analyses, we will relate our results to our framework of teaching, allowing us to provide a thorough account of the key characteristics seen across successful interventions and provide a discussion of the sufficient and necessary conditions – drawing on examples from across studies – for approaches to be deemed successful. It may be that such conditions are the 'ideal' but not necessarily realistic in the mainstream secondary mathematics classroom, in which case we will provide some contextualisation and discussion of how best they may feature in an intervention and what they could look like in practice.

## Registration, data protection and ethics

We do not anticipate data protection or ethical issues, because our dataset will consist of publicly available information and will not include any identifiable personal data. On this basis, the project has received approval from the UCL Institute of Education's Research Ethics Committee (REC1899).

Once finalised, this protocol will be registered and published on the EEF website and pre-registered with the International Platform of Registered Systematic Review and Meta-analysis Protocols (INPLASY).

## Team

All members of the team are affiliated with IOE, UCL's Faculty of Education and Society, University College London.

**Eirini Geraniou (Joint PI)** is a Professor in Mathematics Education at UCL. She has considerable experience in mathematics education research and qualitative research methods, including carrying out systematic reviews of academic and professional literature. She has expertise in training and professionally developing secondary mathematics teachers. She has experience of evaluations such as the IPE for the EEF-funded project of the Mathematical Reasoning Programme. She has co-authored the recent Royal Society report on "Educational Technologies in Mathematics Education", as part of the Mathematical Futures Programme (MFP). She currently serves on the board of the European Society for Research in Mathematics Education. Her most recent work was as PI of the UCL-funded Teachers' Mathematical Digital Competencies project.

**Rachel Marks (Joint PI)** is an Associate Professor in Mathematics Education at UCL. She has considerable expertise in literature reviewing and was the lead researcher for the EEF-funded KS2/3 and EY/KS1 mathematics evidence reviews and led the research review of British research on mathematics education for the British Society for Research into Learning Mathematics (BSRLM). She recently led the Nuffield-funded research study, The prevalence and use of textbooks and curriculum resources in primary mathematics. She has served on the executive of British Society for Research into Learning Mathematics (BSRLM) and was a member of the Independent Commission on Assessment in Primary Education.

**Jeremy Hodgen** will (with the PIs) be a member of the leadership team for the review and will (with Geraniou and Marks) lead on writing the review. He is a Professor of Mathematics Education at UCL and at the Observatory for Mathematical Education at the University of Nottingham. He has expertise in mathematical progression, effective interventions and strategies in mathematics, quantitative and qualitative methods and communicating research findings to practitioners. He has led many funded research projects, including the EEF-funded KS2/3 and EY/KS1 mathematics evidence reviews and contributed to the associated mathematics guidance. He has been a member of the Royal Society's Advisory

Committee on Mathematics Education and is currently a member of the Prime Minister's Expert Advisory Group on Mathematics to 18.

**Nicola Bretscher** will lead on statistical analyses and contribute to the final report. She is a Lecturer in Mathematics Education at UCL. She brings expertise in statistical methods. She was lead statistician for the EEF-funded SMART Spaces Evaluation. She is currently working on the quantitative analysis for the UCL-funded Teachers' Mathematics Digital Competencies project. She has experience of research reviews, including contributing to the Nuffield Foundation-funded report *Values and variables: mathematics education in high-performing countries*.

**Laurie Jacques** will contribute to the coding of studies and the writing of the final report. She has experience of conducting narrative reviews and was a researcher on the recent report, *The mathematics pipeline in England: Patterns, interventions, and excellence* (Noyes et al., 2023). She is a member of the team conducting the concurrent the Secondary Mathematics Practice Review.

**Bohan Liu** will contribute to the coding of studies. She is a doctoral student at UCL IOE.

**Wenfei Du** will contribute to the coding of studies. He is a doctoral student at UCL IOE. He has also recently worked on a systematic literature review with another doctoral student regarding Augmented Reality and Mathematics Education.

## Conflicts of interest

The work described in this protocol is being undertaken by researchers at the IOE, UCL's Faculty of Education and Society and funded by the EEF. The views expressed are those of the authors and not necessarily those of the EEF.

Laurie Jacques is also working as part of the Sheffield Hallam University team leading the EEF Practice Review for KS3 and KS4 Mathematics. She also works as an independent mathematics education consultant for her own business SmartPD Limited.

UCL provides education and other services relating to mathematics education in return for fees or grant income. None of the authors are shareholders or otherwise directly financially benefit (beyond their ongoing employment) from their employers' activity. All authors declare no other conflicts of interest.

## Timeline

Dates	Activity	Staff responsible/leading
Nov-Dec 2023	Plan the project, train team in methods, begin to identify literature, agree on initial themes / topics, inclusion/exclusion criteria & coding system, submit draft protocol for peer review.	Eirini & Rachel
January 2024	Update protocol, responding to reviewers' comments.	Eirini & Rachel & Jeremy
	Team trained in use of EPPI Reviewer	All
	Conduct searches to ascertain all relevant meta-analyses. Unpack (for grey literature) and code, as per this protocol.	Rachel, Laurie, Eirini
February 2024	Conduct and record full searches as per this protocol to produce database for review.	Rachel
February-December 2024	Data extraction, coding (in EPPI Reviewer) and initial analysis.	All
April 2024 – April 2025	Analyse data across topics, produce updated evidence gap map and write the review. <b>Presentation of Initial Findings: mid-June 2024</b>	Eirini & Rachel & Jeremy

	<b><i>SUBMISSION of Updated Evidence Gap Map Review): August 2024</i></b> <b><i>SUBMISSION of Final Report/Review (draft for peer review): 25 April 2024</i></b>	
April - June 2025	Tidy up the data for use by third parties, produce coding manual and ensure all is shareable and accessible on the EPPI platform. <b>Deadline:</b> January 2025	Eirini & Rachel & Jeremy & RAs
July 2025	Publication of final report	

## Protocol Amendments

This protocol was updated in February 2025 to account for minor changes to the methodology and revised procedures for assessing the relevance and quality of studies, as well as the Risk of Bias, individually and across the entire synthesis. These amendments and their location in the protocol are listed below:

Amendment Number	Page	Amendment
1	p. 9, 25	Changed (to avoid contradictory uses of the term quality) “1. A full systematic search (detailed below) for RCTs and high-quality QED studies <u>across</u> all Research Questions and topics. High quality here refers only to the study design of the QED, with all QEDs with a concurrent ‘Business as Usual’ or active control group <u>and</u> a pre-test categorised as of sufficient quality. 2. A systematic search for wider experimental and other quantitative studies (including QEDs not classified as ‘high quality’ such as those without a pre-test).” to “1. A full systematic search (detailed below) for RCTs and QED studies of sufficient quality <u>across</u> all Research Questions and topics. Sufficient quality here refers only to the study design of the QED, with all QEDs with a concurrent ‘Business as Usual’ or active control group <u>and</u> a pre-test categorised as of sufficient quality. 2. A systematic search for wider experimental and other quantitative studies (including QEDs not classified as ‘high quality’ such as those without a pre-test).” Similar change made on p.25.
2	p. 10	We have changed the language of “Modules” to “Topics” making it clear where we build on the work on Modules in the previous reviews and where we are talking about the present review through the use of the term “Topics”.
3	p. 10	Changed “Studies may be coded to one or to multiple topics. It is anticipated that this coding will, in part, support the development of our response to RQ2 as we will be able to assess which topics repeatedly co-occur” to “Studies will be coded to a primary and to a secondary topic. It is anticipated that this coding will, in part, support the development of our response to RQ2 as we will be able to assess which topics repeatedly co-occur”.
4	p. 14	Addition to the study exclusion criteria under “setting”, namely: The intervention setting has a smaller granularity than a class, e.g., we will exclude studies with <15 in each group allocation. [Excluded as a small sample size is one element of deciding that the study is about learning not teaching.]
5	p. 15	Addition to the study exclusion criteria under “intervention type”, namely: The intervention is well-described and could be operationalised by a teacher.
6	pp. 23-4	Addition to “Forward citation searches of full-text includes” to include process developed to identify the highest quality and important studies which may have been missed in previous searches.
7	p. 26	Double-coding changed from a proportion of at least 10% of studies to a threshold of at least 25 studies.
8	p. 26	Clarifications made to the “Assessment of risks of bias” approach and tool, predominantly to account for the use of the more appropriate and rigorous WWC standards for study attrition.
9	p. 27	Addition to “Effect size calculation” to account for studies where the Standard Error or Confidence Intervals are not reported, namely: In cases where an ES is provided but without a SE or CIs, a calculator based on WWC procedures will be used, as advised and supplied by the EEF Toolkit team at the University of Durham. This calculator was based on sample size of students and does not take account of clustering.
10	p. 27	Amendment to effect size reporting to bring methods in line with the reporting of EEF trials and with other high-quality meta-analyses in education, we will report effect sizes using Hedges’ <i>g</i> . Changed “We will correct Cohen’s <i>d</i> for bias in studies with fewer than 50 participants using Hedges’ <i>g</i> .” to “We will correct Cohen’s <i>d</i> for bias using Hedges’ <i>g</i> .”

Amendment Number	Page	Amendment
11	p. 27	Amendment to procedure where multiple outcome measures are reported to clarify that more general and robust measures are preferred. Changed “In cases for which multiple attainment outcomes are specified, we will by preference use the primary attainment outcome. Where multiple relevant attainment outcomes are reported and no primary outcome is specified, we will calculate an aggregated effect size based on calculations of all relevant effect sizes.” To “In cases for which multiple attainment outcomes are specified, we will by preference use the most general and most robust mathematics attainment outcome. Standardised measures are to be preferred over researcher-designed measures. Where multiple relevant attainment outcomes are reported and no primary outcome is specified, we will select an outcome at random.”
12	p. 28	Clarified that potential bias for imbalance at baseline is relevant only for QEDs.
13	p. 30	Addition to “Modular Reporting” to add quality of individual studies assessment process to the overall quality assessment process (see also additional Appendix 4).
14	pp. 30-2	Thresholds for assessing the quality and relevance of overall evidence (including Tables 8 and 9) updated.
15	p. 31	Addition to “Modular Reporting” to add extended and piloted relevance assessment (see also additional Appendix 5) as a measure within itself and as a tool to identify the most important studies within each topic.
16	pp. 33-4	Team information update: updates to title for Eirini Geraniou & institutional affiliation for Jeremy Hodgen
17	pp. 34-5	Timeline updated.
18	p. 53	Appendix 2d (Assessment of risk of bias coding tool) updated to reflect changes in amendment
19	p. 56	Appendix 4 (Quality Assessment for Individual Studies) added as per amendment #13
20	p. 57	Appendix 5 (Relevance Assessment for Individual Studies) added as per amendment #15

## References

- Adkins, M., & Noyes, A. (2016). Reassessing the economic value of advanced level mathematics. *British Educational Research Journal*, 42(1), 93-116. <https://doi.org/10.1002/berj.3219>
- AMSTAR. (2021). *Assessing the methodological quality of systematic reviews: Checklist*. [https://amstar.ca/Amstar\\_Checklist.php](https://amstar.ca/Amstar_Checklist.php)
- Andrews, J., Guyatt, G., Oxman, A., Alderson, P., Dahm, P., Falck-Ytter, Y., et al. (2013a). GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. *Journal of clinical epidemiology*, 66, 719-25.
- Andrews, J., Schunemann, H., Oxman, A., Pottie, K., Meerpohl, J., Coello, P., et al. (2013b). GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. *Journal of clinical epidemiology*, 66, 726-35.
- Baker, S., Gersten, R., & Lee, D.-S. (2002). A Synthesis of Empirical Research on Teaching Mathematics to Low-Achieving Students. *The Elementary School Journal*, 103(1), 51-73. <https://doi.org/10.2307/1002308>
- Boylan, M. (Forthcoming). *Secondary Mathematics Practice Review*. Education Endowment Foundation.
- Cai, J. (Ed.) (2017). *Compendium for Research in Mathematics Education*. Reston, VA: National Council of Teachers of Mathematics.
- Campbell Collaboration. (2016). *Protocol for a systematic review: Targeted school-based interventions for improving reading and mathematics for students with, or at risk of, academic difficulties in Grades K to 6*. <https://campbellcollaboration.org/library/improving-reading-mathematics-academic-difficulties-grade-k-6.html>
- Carbonneau, K. J., Marley, S. C., & Selig, J. P. (2013). A meta-analysis of the efficacy of teaching mathematics with concrete manipulatives. *Journal of Educational Psychology*, 105(2), 380-400. <https://doi.org/10.1037/a0031084>
- Cook, W., Shaw, B., & Morris, S. (2020). *Disadvantage in early secondary school*. Manchester Metropolitan University.
- Deeks, J., Higgins, J., & Altman, D. (2020). Chapter 10: Analysing data and undertaking meta-analyses. In: J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds). *Cochrane Handbook for Systematic Reviews of Interventions version 6.1* (updated September 2020). [www.training.cochrane.org/handbook](http://www.training.cochrane.org/handbook).
- Deloitte. (2012). *Measuring the Economic Benefits of Mathematical Science Research in the UK: Final report to the EPSRC*. Deloitte LLP. <http://www.epsrc.ac.uk/newsevents/pubs/deloitte-measuring-the-economic-benefits-of-mathematical-science-research-in-the-uk/>
- Department for Education (2022). *Opportunity for all: strong schools with great teachers for your child*. [https://assets.publishing.service.gov.uk/media/62416cb5d3bf7f32add7819f/Opportunity\\_for\\_all\\_strong\\_schools\\_with\\_great\\_teachers\\_for\\_your\\_child\\_\\_print\\_version\\_.pdf](https://assets.publishing.service.gov.uk/media/62416cb5d3bf7f32add7819f/Opportunity_for_all_strong_schools_with_great_teachers_for_your_child__print_version_.pdf)
- Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic Interventions for Elementary and Middle School Students With Low Socioeconomic Status: A Systematic Review and Meta-Analysis. *Review of Educational Research*, 87(2), 243-282. <https://doi.org/10.3102/0034654316687036>
- Education Endowment Foundation (2022a). EEF evidence database coding guide: Main data extraction (Version 2.0, March 2022). [https://d2tic4wvo1iusb.cloudfront.net/production/documents/toolkit/MDE\\_CodingGuide\\_V3\\_March2022-1.pdf](https://d2tic4wvo1iusb.cloudfront.net/production/documents/toolkit/MDE_CodingGuide_V3_March2022-1.pdf)
- Education Endowment Foundation (2022b). EEF evidence database coding guide: Effect size data extraction (Version 2.0, March 2022). [https://d2tic4wvo1iusb.cloudfront.net/production/documents/toolkit/ESDE\\_CodingGuide\\_V2\\_March\\_2022-1.pdf](https://d2tic4wvo1iusb.cloudfront.net/production/documents/toolkit/ESDE_CodingGuide_V2_March_2022-1.pdf)
- EEF. (2023). *Protocol for a systematic review: Review protocol template*. EEF.

- Fuchs, L. S., Newman-Gonchar, R., Schumacher, R., Dougherty, B., Bucka, N., Karp, K. S., Woodward, J., Clarke, B., Jordan, N. C., Gersten, R., Jayanthi, M., Keating, B., & Morgan, S. (2021). *Assisting Students Struggling with Mathematics: Intervention in the Elementary Grades (WWC 2021006)*. National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. <http://whatworks.ed.gov/>
- Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, 79(3), 1202-1242.
- Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, 79(3), 1202-1242.
- Gilmore, C., Trundle, R., Bahnmüller, J., & Xenidou-Dervou, I. (2021). How research findings can be used to inform educational practice and what can go wrong: The Ofsted Mathematics Research Review 2021. *Mathematics Teaching*, 278, 35-38. <https://atm.org.uk/write/MediaUploads/Journals/MT278/12.pdf>
- Goos, M., Bennison, A., Quirke, S., O'Meara, N., & Vale, C. (2019). Developing Professional Knowledge and Identities of Non-Specialist Teachers of Mathematics. In D. Potari & O. Chapman (Eds.), *International Handbook of Mathematics Teacher Education: Volume 1* (Second ed.) (pp.211-240). Brill.
- Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650), 924-926.
- Hannes, K., & Lockwood, C. (2012). *Synthesising Qualitative Research: choosing the right approach*. Wiley-Blackwell.
- Hobbs, L., & Torner, G. (Eds.). (2019). *Examining the Phenomenon of "Teaching Out-of-field": International Perspectives on Teaching as a Non-specialist*. Springer.
- Hodgen, J., & Marks, R. (2013). *The Employment Equation: Why our young people need more maths for today's jobs*. The Sutton Trust.
- Hodgen, J., Coe, R., Foster, C., Brown, M., Higgins, S., & Küchemann, D. (2020a). *Low attainment in mathematics: An investigation focusing on Year 9 students in England: Main Report*. UCL Institute of Education. [Hodgen\\_LowAttainersMaths-42015-FinalReport-May2020.pdf \(ucl.ac.uk\)](https://www.ucl.ac.uk/education/research/low-attainment-in-mathematics-2020)
- Hodgen, J., Barclay, N., Foster, C., Gilmore, C., Marks, R. & Simms, V. (2020b) *Early Years and Key Stage 1 Mathematics Teaching: Evidence Review*. Education Endowment Foundation. <https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/early-years-and-key-stage-1-mathematics-teaching>
- Hodgen, J., Foster, C., Marks, R. & Brown, M. (2018) *Improving Mathematics in KS2 and KS3: Evidence Review*. Education Endowment Foundation. <https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/mathematics-in-key-stages-2-and-3>
- Hutchison, J., and Dunford, J., with Treadway, M. (2016). *Divergent pathways: the disadvantage gap, accountability, and the pupil premium*. London: Education Policy Institute. <https://epi.org.uk/wp-content/uploads/2018/01/disadvantage-report.pdf>
- Inglis, M., & Foster, C. (2018). Five decades of mathematics education research. *Journal for Research in Mathematics Education*, 49(4), 462-500.
- Jerrim, J., Greany, T., & Parera, N. (2018). *Educational disadvantage: How does England compare?* Education Policy Institute.
- Kyriacou, C., & Issitt, J. (2008). *What Characterises Effective Teacher-initiated Teacher-pupil Dialogue to Promote Conceptual Understanding in Mathematics Lessons in England in Key Stages 2 and 3: A Systematic Review*. EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
- Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*, 38(5), 635-652. <https://doi.org/10.1080/03054985.2012.734800>



- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE.
- Malouf, D. B., & Taymans, J. M. (2016). Anatomy of an Evidence Base. *Educational Researcher*, 45(8), 454-459.
- Maxwell, B., Stevens, A., Demack, S., Coldwell, M., Wolstenholme, C., Reaney-Wood, S., Stiel, B., & Lortie-Forgues, H. (2021). *Review: EEF Implementation and Process Evaluation (IPE) Quality Pilot*. Education Endowment Foundation.
- Noyes, A., Brignall, C., Jacques, L., Powell, J., & Adkins, M. (2023). *The mathematics pipeline in England: Patterns, interventions and excellence*. University of Nottingham, UK.  
<https://www.nottingham.ac.uk/research/groups/crme/documents/maths-pipeline-report.pdf>
- Noyes, J., Booth, A., Moore, G., Flemming, K., Tunçalp, Ö., & Shakibazadeh, E. (2019). Synthesising quantitative and qualitative evidence to inform guidelines on complex interventions: clarifying the purposes, designs and outlining some methods. *BMJ global health*, 4(Suppl 1) doi:10.1136/bmjgh-2018-000893
- OfSTED [Office for Standards in Education]. (2021). *Research Review Series: Mathematics*.  
[www.gov.uk/government/publications/research-review-series-mathematics](http://www.gov.uk/government/publications/research-review-series-mathematics)
- Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372, n160.  
<https://doi.org/10.1136/bmj.n160>
- Parsons, S., & Bynner, J. (2005). *Does numeracy matter more?* National Research and Development Centre for Adult Literacy and Numeracy (NRDC). <http://nrdc.org.uk/wp-content/uploads/2005/01/Does-numeracy-matter-more.pdf>
- Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, 23(3), 425-438.
- Richardson, M., Tina, I., Barnes, I., Swensson, C., Wilkinson, D., & Golding, J. (2020). *Trends in International Mathematics and Science Study (TIMSS) 2019: National report for England (DFE RR1086)*. Department for Education.
- Schmidt, W. H., Houang, R. T., Sullivan, W. F., & Cogan, L. S. (2022). *When practice meets policy in mathematics education: A 19 country/jurisdiction case study*. OECD Education Working Papers No. 268. OECD.  
<https://doi.org/10.1787/07d0eb7d-en>
- Seidel, T., & Shavelson, R. J. (2007). Teaching Effectiveness Research in the Past Decade: The Role of Theory and Research Design in Disentangling Meta-Analysis Results. *Review of Educational Research*, 77(4), 454-499.  
<https://doi.org/10.3102/0034654307310317>
- Siegler, R. S., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., Thompson, L., & Wray, J. (2010). *Developing effective fractions instruction for kindergarten through 8th grade: A practice guide (NCEE #2010-4039)*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. [whatworks.ed.gov/publications/practiceguides](http://whatworks.ed.gov/publications/practiceguides)
- Simms, V., McKeaveney, C., Sloan, S., & Gilmore, C. (2019). *Interventions to improve mathematical achievement in primary school-aged children: A Systematic Review*. Nuffield Foundation.
- Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Stansfield, C., Van Herwegen, J., Cottingham, S. & Higon, J. (2021a). *What are the characteristics of teacher professional development that increase pupil achievement? Protocol for a systematic review*. London: Education Endowment Foundation.
- Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Van Herwegen, J., & Anders, J. (2021b). *What are the Characteristics of Effective Teacher Professional Development? A Systematic Review and Meta-analysis*. London: Education Endowment Foundation.
- Slavin, R. E., Groff, C., & Lake, C. (2009). Effective Programs in Middle and High School Mathematics: A Best-Evidence Synthesis. *Review of Educational Research*, 79(2), 839-911.

- Star, J. R., Caronongan, P., Foegen, A., Furgeson, J., Keating, B., Larson, M. R., Lyskawa, J., McCallum, W. G., Porath, J., & Zbiek, R. M. (2015). *Teaching strategies for improving algebra knowledge in middle and high school students (NCEE 2014-4333)*. National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. <http://whatworks.ed.gov>
- Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology*, 105(4), 970-987.
- Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What One Hundred Years of Research Says About the Effects of Ability Grouping and Acceleration on K–12 Students' Academic Achievement. *Review of Educational Research*, 86(4), 849-899.
- Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., Bond, M., & Koryakina, A. (2023). *EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis*. EPPI Centre, UCL Social Research Institute, University College London.
- Thomas, J., O'Mara-Eves, A. & Brunton, G. (2014). Using Qualitative Comparative Analysis (QCA) in systematic reviews of complex interventions: a worked example. *Systematic Reviews*, 3(67).
- Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>
- What Works Clearinghouse. (2019). *Teaching Strategies for Improving Algebra Knowledge in Middle and High School Students*. [https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/WWC\\_Algebra\\_PG\\_Revised\\_02022018.pdf](https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/WWC_Algebra_PG_Revised_02022018.pdf)
- White, H., Albers, B., Gaarder, M., Kornør, H., Littell, J., Marshall, Z., Mathew, C., Pigott, T., Snilstveit, B., Waddington, H., & Welch, V. (2020). Guidance for producing a Campbell evidence and gap map. *Campbell Systematic Reviews*, 16(4), e1125. <https://doi.org/https://doi.org/10.1002/cl2.1125>
- White, I. R., & Thomas, J. (2005). Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis. *Clin Trials*, 2(2), 141-151. <https://doi.org/10.1191/1740774505cn081oa>
- Woodward, J., Beckmann, S., Driscoll, M., Franke, M. L., Herzig, P., Jitendra, A., Koedinger, K. R., & Ogbuehi, P. (2012). *Improving mathematical problem solving in grades 4 through 8: A practice guide (NCEE 2012-4055)*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. [http://ies.ed.gov/ncee/wwc/publications\\_reviews.aspx#pubsearch/](http://ies.ed.gov/ncee/wwc/publications_reviews.aspx#pubsearch/)
- Young, J. (2017). Technology-enhanced mathematics instruction: A second-order meta-analysis of 30 years of research. *Educational Research Review*, 22, 19-33. <https://doi.org/https://doi.org/10.1016/j.edurev.2017.07.001>

## Appendix 1 – PISA Participants

Extracted from: <https://www.oecd.org/pisa/aboutpisa/pisa-participants.htm> We include in our review all systems listed (including associated education systems).

<b>Albania*</b>	<b>Egypt*</b>	<b>Latvia</b>	<b>Russian Federation*</b>
<b>Algeria*</b>	<b>El Salvador*</b>	<b>Lebanon*</b>	<b>Rwanda*</b>
<b>Argentina*</b>	<b>Estonia</b>	<b>Liechtenstein*</b>	<b>Saudi Arabia*</b>
<b>Armenia*</b>	<b>Finland</b>	<b>Lithuania*</b>	<b>Scotland*</b>
<b>Australia</b>	<b>France</b>	<b>Luxembourg</b>	<b>Serbia*</b>
<b>Austria</b>	<b>Georgia*</b>	<b>Malaysia*</b>	<b>Singapore*</b>
<b>Azerbaijan*</b>	<b>Germany</b>	<b>Malta*</b>	<b>Slovak Republic</b>
<b>Belarus*</b>	<b>Ghana*</b>	<b>Mauritius*</b>	<b>Slovenia</b>
<b>Belgium</b>	<b>Greece</b>	<b>Mexico</b>	<b>Spain</b>
<b>Bosnia and Herzegovina*</b>	<b>Guatemala*</b>	<b>Moldova*</b>	<b>Sweden</b>
<b>Brazil*</b>	<b>Hungary</b>	<b>Mongolia*</b>	<b>Switzerland</b>
<b>Brunei Darussalam*</b>	<b>Iceland</b>	<b>Montenegro*</b>	<b>Chinese Taipei*</b>
<b>Bulgaria*</b>	<b>India*</b>	<b>Morocco*</b>	<b>Tajikistan* (Dushanbe)</b>
<b>Cambodia*</b>	<b>Indonesia*</b>	<b>Netherlands</b>	<b>Thailand*</b>
<b>Canada</b>	<b>Kurdistan Region (Iraq)*</b>	<b>New Zealand</b>	<b>Trinidad and Tobago*</b>
<b>Chile</b>	<b>Ireland</b>	<b>North Macedonia*</b>	<b>Tunisia*</b>
<b>China (People's Republic of)*</b>	<b>Israel</b>	<b>Norway</b>	<b>Türkiye</b>
<b>Hong Kong (China)*</b>	<b>Italy</b>	<b>Palestinian Authority</b>	<b>Ukraine*</b>
<b>Macao (China)*</b>	<b>Jamaica*</b>	<b>Panama*</b>	<b>United Arab Emirates*</b>
<b>Colombia</b>	<b>Japan</b>	<b>Paraguay*</b>	<b>United Kingdom</b>
<b>Costa Rica</b>	<b>Jordan*</b>	<b>Peru*</b>	<b>United States</b>
<b>Croatia*</b>	<b>Kazakhstan*</b>	<b>Philippines*</b>	<b>Uruguay*</b>
<b>Czechia</b>	<b>Kenya*</b>	<b>Poland</b>	<b>Uzbekistan*</b>
<b>Denmark</b>	<b>Korea</b>	<b>Portugal</b>	<b>Venezuela** (Miranda)*</b>
<b>Dominican Republic*</b>	<b>Kosovo*</b>	<b>Qatar*</b>	<b>Viet Nam*</b>
<b>Ecuador*</b>	<b>Kyrgyzstan*</b>	<b>Romania*</b>	<b>Zambia*</b>

\* OECD Non-Members

\*\* Bolivarian Republic of Venezuela

## Appendix 2 – Data extraction and coding tools

### Appendix 2a – EEF main data extraction tool (version 2) (EEF, 2022a)

Note: This list of codes is not intended to be exhaustive. We anticipate that some additional topic or implementation codes may be required in response to the empirical data. Hence, if the need for new codes arises as we analyse the data, new codes will be added.

Section	Question	Applicable Codes
<b>1. Publication information</b>	1.1 What is the publication type?	Journal article
		Dissertation or thesis
		Technical report
		Book or book chapter
		Conference paper
		Other (please specify)
<b>2. What is the research design and which methods were used?</b>	2.1 What is the intervention name?	Highlight relevant text and code Provide the name of the intervention, programme or approach as given in the report.
	2.2 How is the intervention described?	Highlight relevant text and code Provide a brief summary of the intervention as provided in the report. Please include the rationale for impact on learning if given.
	2.3 What are the intervention objectives?	Highlight relevant text and code Please provide the specific objectives or aims of the intervention, programme or approach as provided in the report
	2.4 Is there more than one treatment group?	Yes (please specify)
		No
		Not specified or N/A
	2.5 How were the participants allocated?	Random allocation (please specify)
		Non-random but matched
		Non-random, not matched prior to treatment
		Unclear
		Not assigned – naturally occurring sample (prospective QED)
		<ul style="list-style-type: none"> <li>Retrospective Quasi Experimental Design (QED)</li> <li>Regression discontinuity (e.g. Policy change)</li> </ul>
	2.6 What was the level of allocation?	Individual
		Class
		School – cluster
		School – multi-site
		Region or district
		Not provided
		Not applicable
	2.7 How realistic was the study?	High ecological validity
		Low ecological validity
		Unclear
<b>3. Where did the study take place?</b>	3.1. Please add information about the location	Specific to the location or place
		Information about the type of location

		No information provided
	3.2. In which country/countries was the study carried out?	Select the country or countries that the study was conducted.
	3.3. What is the educational setting?	Nursery school/preschool
		Primary/elementary school
		Middle school/(Prep)
		Secondary/high school
		Residential/boarding school
		Independent/private school
		Home
		Further education/junior or community college
		Other educational setting (please specify)
		Outdoor adventure setting
		No information provided
<b>4. What is the sample of the study?</b>	4.1. What is the overall sample analysed?	What is the overall sample analysed?
		Other information about the overall sample
	4.2. What is the sex of the students?	Female only
		Male only
		Mixed sex
		No information provided
	4.3. What is the age of the students? (Select ALL that apply)	Select all ages of study participants
	4.4. What is the proportion of low SES/FSM students in the sample?	FSM or low SES student percentage
		Further information about FSM or SES in study sample
		No FSM/SES information provided
<b>5. What was involved in the intervention?</b>	5.1. What type of organisation was responsible for providing the intervention?	School or group of schools
		Charity or voluntary organisation
		University/researcher
		Local education authority or district
		Private or commercial company
		Other (please provide details)
	5.2. Was training for the intervention provided?	Yes
		No
		Unclear/Not specified
	5.3. Who is the focus of the intervention? (Select ALL that apply)	Students
		Teachers
		Teaching assistants
		Other education practitioners
		Non-teaching staff
		Senior management
		Parents
		Other
	5.4. What is the intervention teaching approach?	Large group/class teaching (+6)
		Small group/intensive support (3-5)
		Paired learning
		One to one
		Student alone (self- administered)
		Other (explain in notes)
	5.5. Were any of the following involved in the intervention or approach?	Digital technology (yes/no)
		Parents or community volunteers (yes/no)
	5.6. When did the intervention take place? (Select ALL that apply)	During regular school hours
		Before/after school
		Evenings and/or weekends
		Summer/holiday period
		Other (please specify)
		Unclear/not specified
		Research staff

	5.7. Who was responsible for the teaching at the point of delivery? (Select ALL that apply)	Class teachers Teaching assistants Other school staff External teachers Parents/carers Lay persons/volunteers Peers Digital technology Unclear/not specified
	5.8. What was the duration of the intervention?	
	5.9. What was the frequency of the intervention?	
	5.10. What is the length of intervention sessions?	
	5.11. Are implementation details and/or fidelity details provided?	Qualitative Quantitative No implementation details provided
	5.12. Are any costs for the intervention reported?	Yes (please add details) No
	5.13. Who undertook the outcome evaluation?	The developer A different organisation paid by the developer An organisation commissioned independently to evaluate Unclear/not stated
		Is this an EEF evaluation (yes/no)
<b>6. What kind of primary outcomes are provided?</b>	6.1. What kinds of tests were used?	Standardised test (Please specify) Researcher developed test (Please add details) School-developed test (Please add details) National test or examination (Please specify) International tests (Please specify)
	6.2. Curriculum subjects tested (Select ALL that apply)	Literacy (first language) Mathematics Science Social Studies Arts Languages Other curriculum test
	6.3. In addition to the primary educational attainment outcome, are there other outcomes reported?	Yes No
	6.4. If yes, which other outcomes are reported?	Cognitive outcomes measures (Please specify) Other types of student outcomes (Please specify) Other participants (i.e. not students) outcomes (Please specify)

## Appendix 2b – EEF effect size data extraction tool (version 2) (EEF, 2022b)

Section	Question	Applicable Codes
<b>1. What are the details of the study design?</b>	1.1. What was the study design?	Individual RCT
		Cluster RCT
		Multisite RCT
		Prospective QED
		Retrospective QED
		Interrupted time series QED
		Regression discontinuity with randomisation
		Regression discontinuity – not randomised
		Regression discontinuity – naturally occurring
	1.2. What is the number of schools involved in the study?	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/ clear/ not applicable
	1.3. What is the 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
	1.4. Are details of randomisation provided?	Yes (please specify)
		Not applicable
		No/Unclear
<b>2. How is the sample described?</b>	2.1. What is the sample size for the intervention group?	What is the sample size for the intervention group?
	2.2. What is the sample size for the control group?	What is the sample size for the control group?
	2.3. *What is the sample size for the second intervention group?	*What is the sample size for the second intervention group? (*If there is one)
	2.4. *What is the sample size for the third intervention group?	*What is the sample size for the third intervention group? (*If there is one)
	2.5. Does the study report any group differences at baseline?	Yes
		No/Unclear
	2.6. Is comparability taken into account in the analysis?	Yes
		No
		Unclear or details not provided
	2.7. Is attrition or drop-out reported?	Yes
		No
		Unclear (please add notes)
	2.8. What is the percentage attrition in the treatment group?	What is the percentage attrition in the treatment group?
	2.9. Are the variables used for comparability reported?	Yes
		No
		N/A
	2.10. If yes, which variables are used for comparability?	Educational attainment
		Gender
		Socio-economic status
		Special educational needs
		Other (please specify)
	2.11. What is the total or overall percentage attrition?	What is the total or overall percentage attrition?
		Yes



<b>3. Outcome Details</b> <b>Flow chart of p.23 of the EEF effect size data extraction tool kit to be used to determine the appropriate ES to report. Where an ES needs to be calculated, this will follow the procedure outlined in the protocol and the process will be documented in EPPI as per the coding guide.</b>	2.12. Is clustering accounted for in the analysis?	No
		Unclear
	3.1. Outcomes (CODE AT STUDY LEVEL)	Primary outcome
		Secondary outcome
		SES/FSM outcome
	3.2. Are descriptive statistics reported for the primary outcome?	Yes
		No
		Unclear
	3.2.1. 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?
	3.2.2. If yes, please add for the control 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?
	3.2.3. If yes, please add for the second intervention group (if needed)	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?
	3.2.4. If yes, please add for the second control group (if needed)	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?
	3.2.5. If yes, please add for the third intervention group (if needed)	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?
	3.2.6. If yes, please add for the third control group (if needed)	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?
	3.3. Is there follow-up data?	Yes



<b>3. Outcome details (cont.)</b> <b>Flow chart of p.33 of the EEF effect size data extraction tool kit to be used to determine the appropriate classification of an outcome.</b>		No
	3.4.1. Sample (select one from this group)	Sample: All
		Sample: Exceptional
		Sample: High achievers
		Sample: Average
		Sample: Low achievers
	3.4.2. 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
	3.4.3. Effect size calculation (select one from this group)	Post-test unadjusted
		Post-test adjusted for baseline attainment
		Post-test adjusted for baseline attainment AND clustering
		Pre-post gain
	3.4.4. Outcome type (select all that apply)	Toolkit primary outcome
		Reading primary outcome
		Writing and spelling primary outcome
		Mathematics primary outcome
		Science primary outcome
	3.4.5. Toolkit strand(s) (select at least one Toolkit strand)	Other outcome
		Toolkit: Arts participation
		Toolkit: Aspiration interventions
		Toolkit: Behaviour interventions
		Toolkit: Built environment
		Toolkit: Collaborative learning
		Toolkit: Early literacy approaches
		Toolkit: Early numeracy approaches
		Toolkit: Early years intervention
		Toolkit: Extending school time
		Toolkit: Feedback
		Toolkit: Homework
		Toolkit: Individualised instruction
		Toolkit: Mastery learning
		Toolkit: Metacognition and self-regulation
		Toolkit: Mentoring
		Toolkit: One to one tuition
		Toolkit: Oral language interventions
		Toolkit: Outdoor adventure learning
		Toolkit: Parental engagement
		Toolkit: Peer tutoring
		Toolkit: Performance pay
		Toolkit: Phonics
		Toolkit: Reading comprehension strategies
		Toolkit: Reducing class size
		Toolkit: Repeating a year
		Toolkit: School uniform
		Toolkit: Setting or streaming
		Toolkit: Small group tuition
		Toolkit: Social and emotional learning
		Toolkit: Sports participation
		Toolkit: Summer schools
		Toolkit: Teaching assistants
		Toolkit: Within-class attainment grouping
	3.5.1. Comparison	With active control
		With business as usual
		With no equivalent teaching
		Literacy: reading comprehension

	3.5.2. Intervention outcome measure	Literacy: decoding/phonics
		Literacy: spelling
		Literacy: reading other
		Literacy: speaking and listening/oral language
		Literacy: writing
		Mathematics
		Science
		Social studies
		Arts
		Languages
		Curriculum: other
		Combined subjects
		Cognitive: reasoning
		Cognitive: other

## Appendix 2c – Topic and intervention evaluation coding tool

Note: This is an indicative list that will be developed through our ongoing analysis (particularly Section 5 and 6: Topic Mapping, and. Intervention features, components and characteristics.

Section	Question	Applicable Codes
1. Study design	1.1 What is the study design?	RCT
		High-quality QED
		Lower-quality QED (e.g., natural experiment; difference-in-difference; regression-discontinuity; interrupted time series; matched design; cross-over)
		Pre-post-test design or before/after design
		Correlational study
		Cohort study
		Other quantitative design (please specify)
2. Outcomes	2.1 What is the primary outcome measure?	Free text / Highlight the corresponding information in the source
	2.2 What is the headline finding of the study?	Free text / Highlight the corresponding information in the source
3. Intervention	3.1 What is the intervention name?	Free text / Highlight the corresponding information in the source
	3.2 How is the intervention described?	Free text / Highlight the corresponding information in the source
	3.3 Who delivered the intervention?	Class teacher
		TA
		Other member of school staff (please specify)
		External (e.g. researcher) (please specify)
	3.4 Is the intervention replicable as described?	Yes with fidelity
		Yes with interpretation
		No
	3.5 Are inherent features in the implementation of the intervention discussed (e.g. teacher PD)?	Yes (please specify all features)
		No
4. Mathematical topic / aims	4.1 Which topic area(s) does the study cover?	General
		Number
		Algebra
		Ratio, proportion and rates of change

5. Topic mapping	4.2 Does the study explicitly seek to develop learners' skills in working mathematically?	Geometry and measures
		Probability
		Statistics
		Fluency
		Reasoning
		Problem solving
	5.1 Effective strategies	Assessment & feedback
		Cooperative learning (including group work)
		Cognitive science-informed approaches (e.g., spaced learning, interleaving, WM)
		Explicit teaching /Direct instruction (including explicit, whole class, structured, traditional)
		Discussion, talk & language
		Heuristics
		Integrative approaches
		Mastery
		Peer-tutoring (including cross-age tutoring)
		Student-centred (including modern, inquiry/problem-based teaching/posing, individualised approaches)
		Systematic/programmed instruction, curriculum design, data driven instruction
		Thinking skills, metacognition and self-regulation
	5.2 Wider school-level strategies	Family engagement
		Grouping students
		Homework
		Tutoring by adults
		Other (after-school etc. – specify)
	5.3 Resources, contexts and representations	Calculators
		Manipulatives
		Real-life contexts (including modelling)
		Representations (including manipulatives)
		Tasks
		Technology (including CAI & ITS)
		Textbooks
	5.4 Transition	KS2 to KS3
		KS4 to KS5 (A Level)
		KS4 to KS5 (non-A-Level mathematics courses)
	5.5 Teachers	Leadership

		Professional Development
		Specialist teachers
		Non-specialist teachers
		Other teacher outcomes (specify)
	5.6 Other	Anxiety
		Attitudes and motivation
		Behaviour
		Diagnosing difficulties
		Different attainment levels
		Metacognition
		Misconceptions
6. Intervention features, components and characteristics	6.1 Pedagogic features, components and characteristics	To be agreed
	6.2 Implementation features, components and characteristics	To be agreed

## Appendix 2d – Assessment of risk of bias coding tool

Section	Question	Applicable Codes
1. Attrition	1.1 Is any outcome data missing for any reason by allocated group?	Meets WWC Conservative Threshold
		Meets WWC Liberal Threshold
		Does not meet WWC Liberal Threshold
		Extremely high (beyond WWC scale >58%)
		Not reported
2. Non-compliance	2.1 Is compliance evaluated and reported?	Quantitative evidence reported
		Only qualitative evidence reported
		Not reported
3. Pre-specification	3.1 Was the analysis pre-specified?	0 – no
		1 – Yes
		Not reported
4. Analysis level	4.1 Was the experiment analysed at the same level as it was randomised?	1 – Yes
		Not reported
5. Units randomised	5.1 What was the number of units randomised?	Free text – insert sample size
6. Baseline equivalence (for QED studies (of medium or higher relevance))	6.1 Are members of the treatment group the same as members of the comparison group before the study began?	0 – no
		1 – Yes
		Not reported

## Appendix 3 – Pseudo code for meta-analysis

```
## Use metafor package in R

## https://www.metafor-project.org/doku.php/metafor

library(metafor)

## Load data for topic (here 'calculators')

Data <- read.csv("Metadata.csv", header = TRUE, sep=",")

Data <- subset(Data, calculators=="Y")

## Fit random effects model using rma function

## Using restricted maximum likelihood estimator (REML)

res <- rma(yi=effectsize, vi=var, data=Data, slab=paste(author, year, sep=" ", method="REML")

## Create forest plot

forest(res, xlab="Effect size (d)")

## Check for influence of potential outliers using influence and leave1out functions

inf <- influence(res)

l1o <- leave1out(res)

## Create contour-enhanced funnel plot

funnel(res, level=c(90, 95, 99))

## Conduct regression test for funnel plot asymmetry

regtest(res)

## Check for imbalance using trim & fill method and create related funnel plot

taf <- trimfill(res)

funnel(taf, level=c(90, 95, 99))

## Conduct selection model analysis using Citkowitz & Vivea (2017) approach
```

```
Sel <- selmodel(res, type = "beta")
```

```
## Meta-regression to assess the relative effect of different moderators
```

```
Res_mod <- rma(yi=effectsize, vi=var, mods=~facotr(moderators), data=Data, slab=paste(author, year, sep=", "),  
method="REML")
```



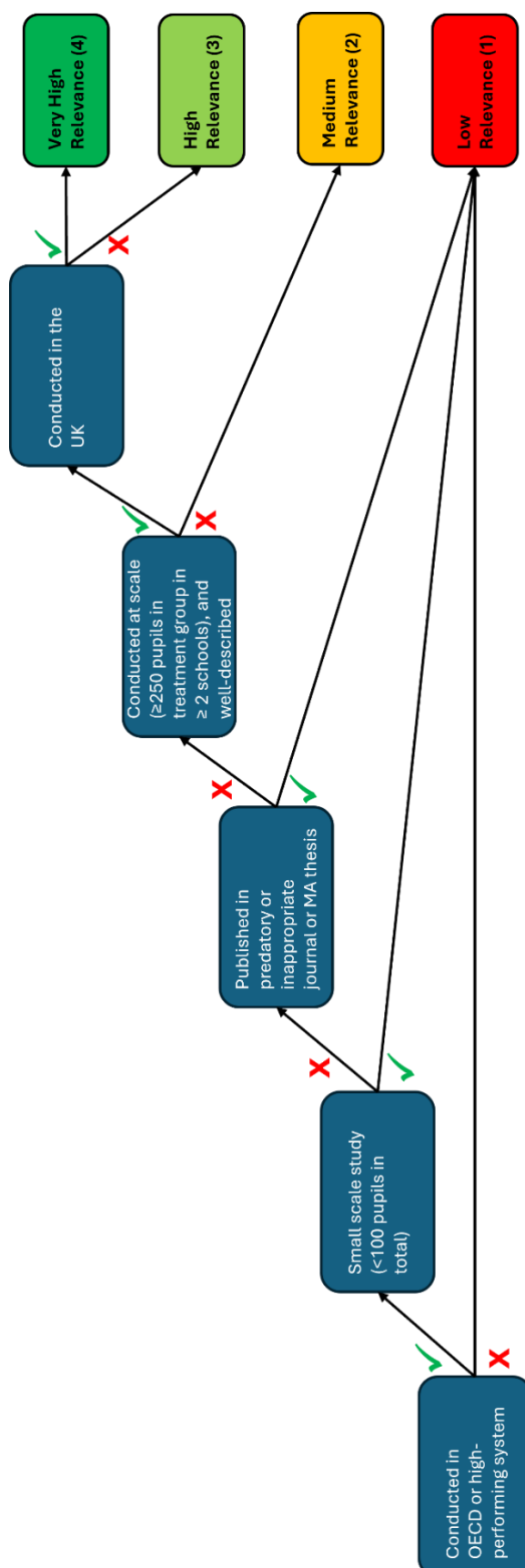
## Appendix 4 – Quality Assessment for Individual Studies

This quality of evidence categorisation is to be used to classify the quality of individual studies, then these classifications will be collated in order to judge the quality of the evidence as a whole by intervention topic and across the entire synthesis.

Quality Category	Study Requirements
Excellent quality [5]	RCT, very large scale: 500 (or more) pupils in the study or 25 (or more) clusters (classes or schools), Meets conservative Attrition Level, Pre-specified design implemented (including power calculations), Analysis good (clustering accounted for, analysed at same level as randomised), Standardised test, Compliance data collected & reported (without problems), no significant threats to validity identified.
Very high quality [4]	RCT, very large: 500 (or more) pupils in the study or 25 (or more) clusters (classes or schools), Meets conservative Attrition Level, Analysis good (clustering accounted for, analysed at same level as randomised), standardised test, no significant threats to validity identified.
High quality [3]	RCT, large scale: 250 (or more) pupils in the study or 10 (or more) clusters (classes or schools), Meets liberal Attrition Level, Analysis good (clustering accounted for, analysed at same level as randomised), no additional significant threats to validity identified. OR QED, large scale: 250 (or more) pupils in the study or 10 (or more) clusters (classes or schools), Meets conservative attrition, Analysis good (clustering accounted for, analysed at same level as randomised), Standardised test, Compliance data collected & reported (without problems), no significant threats to validity identified.
Medium quality [2]	RCT or QED, large: 250 (or more) pupils in the study or 10 (or more) clusters (classes or schools), attrition not extremely high [beyond WWC scale >58%, Analysis good (clustering, units analysed)]
Low quality [1]	RCT or QED, medium scale: 100 (or more) pupils in the study, attrition not extremely high [beyond WWC scale >58%].
Very low quality [0]	All other included studies

## Appendix 5 – Relevance Assessment for Individual Studies

This flow chart will be used by the team to assess the relevance of each study to mathematics teaching in Key Stage 3 and Key Stage 4 in England. Relevance judgements will be used to identify the most important studies within the topic as well as supporting the reader in understanding the overall shape of the literature base.



You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

To view this licence, visit <https://nationalarchives.gov.uk/doc/open-government-licence/version/3> or email: [psi@nationalarchives.gsi.gov.uk](mailto:psi@nationalarchives.gsi.gov.uk)


Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at <https://educationendowmentfoundation.org.uk>



The Education Endowment Foundation  
5th Floor, Millbank Tower  
21–24 Millbank  
London  
SW1P 4QP

<https://educationendowmentfoundation.org.uk>

 [@EducEndowFoundn](https://twitter.com/EducEndowFoundn)

 [Facebook.com/EducEndowFoundn](https://Facebook.com/EducEndowFoundn)