Evaluation summary

Project title	Effectiveness trial of Mathematical Reasoning
Developer *(Institution)*	University of Oxford
Evaluator *(Institution)*	National Foundation for Educational Research (NFER)
Principal investigator(s)	Helen Poet
Protocol author(s)	Lillian Flemons, Andrew Smith, Chris Morton, Helen Poet
Trial design	Two-arm cluster randomised controlled trial with random allocation at the school level.
Trial type	Effectiveness
Pupil age range and Key stage	Year 2
Number of schools *(at design stage)*	240
Number of pupils *(at design stage)*	6,168
Primary outcome measure and source	Progress Test in Maths 7 (PTM7) test administered by teachers at baseline and by NFER test administrators at endpoint
Secondary outcome measure and source[1]	GL Assessment Progress Test in Maths (PTM7) 'process' categories (subscales): (i) fluency in facts and procedures, (ii) fluency in conceptual understanding, (iii) problem-solving, and (iv) mathematical reasoning.

[1]For RQ2.1. The other secondary RQs use the same outcome as primary RQ1.

Protocol version history

Version	Date	Reason for revision
1.0 [original]	03 July 2024	N/A
1.1	30 June 2025	New staff member; modified approach to collecting training completion data; minor change to survey content; impact analysis changes agreed when writing SAP: removal of sensitivity check that excludes 2 PTM6 items; changes to compliance definitions; addition of models outside compliance analysis looking at impact of training and session completion; addition of MNAR check to missing data analysis.

Contents

Tables

Figures

Study rationale and background

Mathematical Reasoning (MR) is a programme for pupils in Year 2 that aims to improve mathematical attainment for all pupils by developing their understanding of the logical principles underlying mathematics. The programme was developed by Professor Terezinha Nunes and Professor Peter Bryant at the University of Oxford in response to their own research that found these two abilities at age 8-9 years predicted KS2 and KS3 mathematical attainment (Nunes *et al.*, 2012). This project also built on research that found additive reasoning and logical abilities to be predictive of 6-year-olds' mathematical attainment 12 to 16 months later (Ching and Nunes, 2017; Nunes *et al.*, 2007). Early number sense has likewise been found to predict later mathematics achievement (Jordan, Devlin and Botello, 2022). The programme is based on the KS1 National Curriculum and introduces no new content. MR is not currently commercially available.

The programme focuses on quantitative reasoning and number sense and replaces one lesson per week for 12 weeks with a programme session. Each session involves both whole-class teacher-led time and differentiated group time, alternating between tailored teacher support and time spent on programme-specific computer games to embed learning. Quantitative reasoning and number sense are distinct but related skills, and both have a role to play in learning arithmetic. The developer team defines quantitative reasoning as the ability to reason about quantities and relations between quantities, with or without numbers. Number sense is defined as the ability to reason about relations between numbers using the four operations. Within this latter domain, the programme focuses specifically on additive composition (i.e. any number can be seen as the sum of two other numbers) and the inverse relation between addition and subtraction (i.e. when a certain number is both added and taken away the original amount remains the same) (Ching *et al.*, 2020). While the KS1 curriculum tends to teach arithmetic operations before applying these concepts to problem solving, the MR programme seeks to provide pupils with the quantitative reasoning and number sense first as a foundation for problem solving skills. The programme treats quantitative reasoning and number sense as complementary but separate strands of learning, reflecting previous research that has found that a child's understanding of quantities and their understanding of numbers are not always connected (Ching *et al.*, 2020).

Teachers and TAs are trained to deliver the approach through e-learning and professional development support. The programme aims to improve teacher and teaching assistant (TA) pedagogical knowledge around mathematical and numerical reasoning and to increase understanding of the importance of teaching these concepts and skills from a young age. While the core purpose of the CPD is to train teachers and TAs to effectively deliver the MR programme, it is also intended to empower teachers and TAs to apply their learning from the programme to other areas of their work and to share it with their colleagues. In this way, pupils not directly involved in the programme may still benefit from the increased understanding and improved practice of teaching staff in these areas. The CPD element is delivered as an online training programme made up of five core modules and four webinars accompanied by tailored implementation support from trained professionals.

Previous trials

The MR programme has been the subject of two randomised controlled trials (RCT) commissioned by the Education Endowment Foundation (EEF), both with a high-security rating.

The efficacy trial (Worth *et al.*, 2015) was a three-arm RCT conducted by the National Foundation for Educational Research (NFER), with 17 out of 55 schools allocated to the MR group.[2] In that efficacy trial, the Oxford University team trained teachers directly through a one-day in-person training session, with one follow-up visit to each participating school to observe delivery and provide personalised feedback. The trial found that the programme achieved a positive impact on pupils' numeracy abilities, equating to, on average, three months' progress (effect size of 0.20), compared to pupils who had not received the programme. A slightly smaller impact of 2 months' progress was found for pupils eligible for Free School Meals (FSM) (effect size of 0.14).

The subsequent effectiveness trial (Stokes *et al.*, 2018), carried out by the National Institute of Economic and Social Research (NIESR), was an RCT involving 160 schools. A train-the-trainer model was employed for this trial. The National Centre for Excellence in the Teaching of Mathematics (NCETM) helped to develop the training model, which was delivered through by the national network of 'Maths Hubs'.[3] The trial found a smaller impact of one month's progress for all pupils and pupils eligible for FSM specifically (effect size of 0.08 and 0.09 respectively); however, these results were not statistically significant.[4]

In response to the limited impact observed in the effectiveness compared to the efficacy trial, the Oxford University team sought to improve the fidelity of delivery at a larger scale by developing a fully asynchronous online training course created by the programme developer to remove the intermediary effect of the train-the-trainer model. Additional tailored support over the course of the implementation period is provided to schools by Teacher Leaders (TLs) trained by the Oxford team via webinars, email and an online forum. Other elements of the programme remained unchanged from previous iterations of the MR programme, including the structure and content of the sessions (including both teacher-led and differentiated group work) and the use of computer games. In advance of this trial, a pilot evaluation of the new training model was carried out by the Institute of Education (IOE) at University College London (UCL) in 2023, with the final report forthcoming. The pilot study aimed to recruit 32 schools and sought to assess how effective the new training model was for preparing to teachers to deliver the programme, as well as the feasibility of implementing this model and its scalability. Interviews, observations and surveys were carried out to inform this assessment. More information about the version of the MR programme being evaluated can be found in the Intervention section below.

[2] At this point the programme was named Mathematics and Reasoning.

[3] Maths Hubs are partnerships of schools focused on maths education.

[4] Confidence intervals were (-0.03, 0.18) and (-0.07, 0.25), respectively.

The current trial

Given the positive impact identified in the efficacy trial, the EEF is interested in determining how to implement the MR programme at a larger scale most effectively. The interim findings from a recent pilot study[5] indicated positive results for the acceptability and feasibility of the new training model and its effectiveness in preparing teachers to deliver the programme. As a result, the EEF is commissioning this second effectiveness trial to understand whether this new training model may better retain the scale of impact seen at the efficacy stage by enabling direct contact with the Oxford University team's teaching material. A summary of the differences in both the programme and evaluation for each of the trials can be found in Appendix A.

The integrated evaluation includes both impact and implementation and process evaluation (IPE) components. The primary focus of the impact evaluation will be to estimate the impact of the programme on short-term pupil mathematical attainment outcomes (as per the Theory of Change (ToC)). It will also investigate the effects of pupil prior attainment and dosage and compliance, all of which are hypothesised to be potential moderators of outcomes. The design of the impact evaluation is broadly congruent with the previous effectiveness trial to allow for some comparison of findings and as the potential basis for inference about the revised training model. As the sample size for the efficacy trial meant that inconclusive results were found for pupils eligible for FSM, this trial aims to provide a more accurate assessment of what the effect for this particular sub-group may be.

The IPE will particularly focus on the fidelity of implementation and how this relates to the nature and quality of the training and support provided. We will also be looking to understand better the pupil grouping practices implemented as part of the programme delivery by teachers. This is an important part of the programme's differentiated teaching model, which is intended to be implemented flexibly each week, so we intend to explore actual practice given the evidence around ability grouping for lower-ability children (Henry, 2015; Johnston & Wildy, 2016; Parsons & Hallam, 2014),. In-depth case studies involving observations, pupil focus groups and interviews with staff members at two-time points will allow us to explore implementation factors like this in detail. The use of computer games, which was seen to vary significantly in the previous trials, will likewise be a point of focus.

Intervention

The MR programme will be implemented for the purpose of the effectiveness trial between December 2024 and April 2025.

A detailed description of the programme in the context of the TIDieR checklist is presented below. A summary of the training and preparation carried out by TLs can be found in Appendix D.

5 Pilot report publication forthcoming

Why: Rationale, theory and/or goal of essential elements of the programme

The aim of the MR programme[6] is to improve mathematical attainment by developing pupils' understanding of the logical principles underlying mathematics, primarily:

1. Quantitative reasoning: the ability to reason about quantities and relations between quantities (with or without numbers).
2. Number sense: the ability to reason about relations between numbers using the four operations, specifically focusing on additive composition and the inverse relation between addition and subtraction.

The programme does not introduce any new subject content outside of the national curriculum but instead seeks to improve reasoning and understanding of existing concepts through a teaching approach that emphasises quantitative reasoning and number sense as the foundation for problem-solving and arithmetic. The programme promotes discussion and the use of manipulatives (by both the teacher and the pupils) to support mathematical thinking.

The causal logic of the programme (see Figure 1) is based on previous research by the developers that found an association between pupils' quantitative reasoning and arithmetic abilities and their subsequent mathematical attainment, even from early primary school (Nunes *et al.*, 2012; Ching and Nunes, 2017; Nunes *et al.*, 2007). The programme also seeks to support the achievement of these outcomes by increasing pupil confidence and enjoyment around maths, as there is evidence to suggest that these factors may be predictors of subsequent mathematical attainment (Çiftçi and Yildiz, 2019; (Putwain *et al.*, 2018).

Who: Recipients of the programme

The programme recipients are Year 2 pupils in state schools in England. The programme is delivered to the whole class, with reasonable adjustments made for pupils with special educational needs or disabilities (SEND) where necessary.

What: Materials

Training for schools

All teachers and TAs are given access to an online training course comprising nine modules. The first five modules prepare the teacher/TA pair to implement the programme, while the subsequent four aim to show how mathematical reasoning can be used in teaching other mathematics topics in primary schools. Each module includes brief video lectures from a member of the Oxford team explaining key ideas before the programme and implementation guidelines, as well as downloadable research briefs and presentation slides with further details about the programme and the theory and evidence behind it. The course is self-guided and includes interactive elements, opportunities for reflection and videos demonstrating how different sessions should be delivered. Teachers and TAs can interact with their TL and other schools in their cohort (see below) via a chat function embedded in the course.

[6] https://reasoningfirst.org.uk/programmes/mathematical-reasoning-year-2/

Programme delivery

Each school receives digital presentation slides for each of the 12 sessions, as well as a teaching handbook for the programme. This handbook includes a detailed unit plan for each session and instructions for each whole-class and group activity, as well as a glossary. A document with Frequency Asked Questions (FAQs) is also provided.

As part of the programme, each pupil receives a Pupil Workbook, which includes written activities and extension worksheets for the teacher-led components, as well as cut-out shapes to be used as manipulatives.

Schools are also provided with (paper) worksheets with supplementary games that pupils can complete before starting on the computer games if they are not yet at the ability level required for the first online game.

Schools are expected to provide the necessary IT equipment, including a screen for presenting the slides during the whole-class component of the sessions. In addition, schools are expected to provide additional manipulatives for pupil use, such as counters, blocks and coins.

Computer games

In the second part of the MR session, around half of the class is allocated to play computer games. These games were created by the University of Oxford for this programme as a tool for pupils to practice their MR skills. Children access the games through a website. They are assigned an individual log-in so that progress can be tracked and achievements rewarded. Schools are given access to and instructions for the computer games website, which includes bonus games and certificates of achievement that can be downloaded and printed separately. Each pupil receives a Pupil Record Sheet to track the games they play. Schools must provide access to computers or tablets for pupils to play computer games during the group component of the sessions. There is no minimum number of computers or tablets required – this is up to the capacity and discretion of each school.

What: Procedures, activities and/or processes

Training for schools

The teacher and a TA for each participating class is expected to complete the five core online modules before starting programme delivery. The first four modules cover the theory and rationale and provide an overview of the programme. The fifth module provides practical guidance for implementing the programme as well as highlighting the importance of fidelity and the kinds of adaptations that would be acceptable. The TL provides each teacher/TA pair with access to the fifth module once they judge them to have satisfactorily completed modules 1 to 4 based on their responses to the activities in each of the modules. These activities are not intended to be a knowledge assessment but instead require the respondents to share reflections on what they have learnt.

The training course includes an additional four online modules that offer teachers and TAs the opportunity to learn more about how mathematical reasoning can be promoted when teaching other topics in the maths curriculum (such as fractions and diagrams), including for older age

groups. Only the last of these modules (module 9) is compulsory. It covers how participants can use what they have learnt in other areas of their teaching practice and share this knowledge with other teachers in the school. Teachers and TAs are expected to complete this module ahead of the final webinar (see below).

The TLs monitor the online course and provide support via the discussion forums and chat function in the online training course and by email. Each TL supports a separate cohort of around 20 schools. There are no criteria for allocating schools to particular cohorts or TLs. Online interaction is restricted to within each cohort.

Each teacher/TA pair receives one set of log-in details, and the teacher and TA are encouraged to complete the training course together to promote knowledge sharing and discussion.

In addition to the online training, teachers and TAs are expected to attend three live webinars delivered by their TL over the delivery period. Each TL delivers a webinar attended by their own cohort only.[7] The materials for these webinars are designed by the Oxford University team but with a focus on participant-led discussion. The webinars seek to:

- provide support to teachers and TAs in tackling practical challenges that may arise at any phase of the programme,
- create a community of practice for reflective thinking and peer learning, with sharing between schools,
- support integration of learning into practice beyond the programme.

The first webinar provides the theoretical background to the programme and outlines each of its key components. The second webinar focuses on knowledge-sharing between schools and introduces some of the activities that feature in later sessions. The third webinar looks at the sustainability and future use of the programme in the school and how it can apply to other year groups and/or areas of the curriculum.

Finally, TLs also provide schools with practical and administrative support.

Programme delivery

The programme consists of 12 units delivered by the teacher, with TA support, across 12-15 sessions. Each session comprises a whole-class component and a group component. For the group component, the teacher divides the class into two groups: L1 and L2. L1 consists of pupils for whom the teacher feels additional support could be beneficial, while L2 is for pupils perceived by the teacher to be ready for further learning. These groupings are intended to be flexible according to perceived pupil needs in response to each topic covered on a session-

[7] Teachers are expected to participate in the webinars led by the TL to whose cohort they were assigned, but if they cannot attend at the designated time(s) they are encouraged to attend a webinar led by one of the other TLs instead.

by-session basis. The groups alternate between teacher-led activities and playing computer games, according to the allocation provided for each unit in the handbook.

The programme consists of two conceptual strands (number sense and quantitative reasoning). The activities that explore each of these are intended to be interleaved such that the pupils' skills in each develop in tandem. The number sense strand focuses on the additive composition of numbers, place value and inverse relations between addition and subtraction. Pupils are encouraged to think about the changes caused by performing and undoing actions, not just by counting forward or backwards. The quantitative reasoning strand focuses on the different relations that can be established between quantities and asks pupils to visualise relational scenarios using manipulatives. The programme starts by working with smaller numbers before progressing to larger numbers.

The first part of each session involves the whole-class component led by the teacher and is expected to last around 40 minutes. In each session, the teacher introduces the (new) concept, often with manipulatives and an animated presentation, and presents the class with story problems to solve. The pupils each write an answer in their Pupil Workbook and discuss it in pairs, using manipulatives to inform their working out by enacting the story problems. The teacher then talks through the work, asking the pupils to explain the thinking and processes behind their answers. There are extension exercises at the back of the Workbook for pupils who finish an activity early to complete while waiting for the rest of the class. During this session, the TA makes sure pupils have the materials they need and are answering the questions in their workbooks. The TA also supports pupils to turn to the extension activities and/or to make reasonable adjustments to the activities according to pupil needs.

The remainder of each session is spent in group activities. One group works with the teacher, who provides extra support or pre-teaching for L1 and extension opportunities for L2. The handbook provides detailed instructions for the activities to be carried out in the teacher-led group component for each unit and whether the activities should be with L1 or L2 in each case.

The other group plays computer games, which provide pupils with the opportunity to practice the concepts taught in the whole-class session to consolidate their learning. The games are divided into units that reflect the structure and content of the lessons. However, pupils can work through the units at their own pace. Pupils record the games they have completed on their individual Child Record Sheets. The TA supervises the group playing the computer games, helping pupils to log on and complete their Record Sheets, showing them how different games work and encouraging them to play a variety of games. Pupils who achieve 100% on a game receive a certificate (shared with the pupil at the TA/teacher's discretion) and the opportunity to play a short bonus game. The computer games can also be accessed from home, and schools will be provided with guidance on how to facilitate this.

There are supplementary activities for any pupils who are not yet at the ability level required to play the first set of additive composition computer games.

Who: Programme providers/implementers

The developer team at the University of Oxford provides all the material and trains the TLs.

The TLs support teachers in completing the training and delivering the programme. They are specialist teachers trained in the programme by the developers and have extensive experience in training other teachers to implement programmes and/or delivered the programme as a teacher at least twice in the past.

Nominated Year 2 class teachers are responsible for delivering the teacher-led components of the programme (whole-class and group).

Nominated TAs are responsible for supporting the teachers in delivering the programme, including the whole-class components, and monitoring the use of computer games during the group components of the sessions.

How: Mode of delivery

Teacher and TA training is online via a training course, webinars and a forum.

The programme is delivered in person, and children play games online for part of each session.

See the 'Materials' and 'Procedures' sections above for further details.

Where: Location of the programme

The programme is intended to be delivered in the regular classrooms of participating schools. It is recommended that pupils sit on the carpet for the whole-class component, but this is up to the teacher's discretion, as long as the pupils are situated so as to be focused on the teacher and questions on the screen at the front and are able to engage in discussions with the whole class.

Schools can be located anywhere in England for the purpose of the trial.

When and how much: Duration & dosage

Training for schools

Modules 1-5 can be completed at times convenient to the teachers within a designated period to create a learning community of teachers and to ensure they are completed in good time. Headteachers are asked to release teachers and TAs for one and a half days to complete modules 1 to 5 plus module 9 and to set the pupils up on the computer games. Ideally this time is allocated in half-day blocks, rather than short sessions across a longer period.

The first webinar takes place once schools have completed the training but before they start delivery, the second takes place two to three weeks into the delivery period, and the third shortly before the end of the delivery period.

Programme delivery

The programme consists of 12 units delivered weekly across 12-15 weeks, depending on the number of sessions the teacher spends on each unit (each unit does not necessarily equate to one session). Each session should last approximately one hour (or the length of a normal lesson) and take place during normal maths lesson time. Approximately 40 minutes of each session should be spent on the whole-class component and 20 minutes on the group activities.

If the teacher is not able to complete the activities from a whole class lesson within 40 minutes, they should still move on to the group component of the session and start the whole-class session in the next lesson at the point they stopped.

While in practice the programme can be delivered in classrooms at any point after completing the training, for the purpose of the trial, it will be delivered across the autumn and spring terms (following randomisation, baseline testing, and training completion). Endpoint testing will occur in the summer term.

Tailoring: Adaptation of the programme

High fidelity to the course structure, material, and content is required. Teachers are told not to skip any of the content and to continue the next session where they left off if a whole unit cannot be covered in a single session.

Duration and timing of different aspects of the programme are more flexible, including when teachers and TAs complete the training, how long sessions last, how long is spent on each component of the session, and how many weeks the programme lasts. The proportion of the class assigned to each of the groups (L1/L2) is also flexible, depending on perceived pupil need.

For this trial, Oxford plans to provide schools with guidance asking teachers to encourage the use of computer games at home. The guidance will also include how to support disadvantaged pupils (e.g., encouraging schools to provide access at school for those without access at home where possible).

Teachers are encouraged to make any necessary adjustments for pupils with SEND, as they would in a normal lesson. They should also tailor the pace and approach of each teacher-led group component to the needs of that particular group.

Examples of broadly acceptable adaptations include (also see 'How well (planned)' section below):

- doing two sessions per week,
- skipping two weeks when school has lots of other planned activities, and starting again or repeating the last session,
- allowing the pupils to play the games during their free time.

Examples of more significant adaptations that would broadly be considered unacceptable include:

- changing the order in which the programme content is delivered,
- starting from session 5 because the early sessions seem too easy,
- replacing inverse relation tasks with practice in number facts,
- skipping the whole class discussion to catch up when behind or to go faster during the session,

- skipping the group component of a session (which would mean skipping the computer games) or alternating between whole-class and group sessions,
- using larger numbers in the activities right from the beginning because the pupils already know how to count to 100.

How well (planned): Strategies to maximise effective implementation

Teachers are provided with a list of 'keys to success' to optimise for effective implementation (see Appendix C). These include what to do should the teacher be unable to deliver all the content for each session within the space of a single lesson – something which had emerged as an issue in the previous trials.

A section of the online training course explicitly emphasizes the importance of fidelity and the kinds of adaptations to the programme that may or may not be acceptable.

TLs are provided with training (see Appendix D) to carry out their role, including supporting schools with any queries or challenges in delivering the programme. Schools are also provided with numerous different opportunities and forums for raising questions or concerns with their TL, including online forums, webinars and email.

Both teachers and TAs are asked to complete the training together, with the aim of promoting discussion about the programme and consequently reducing the risk of misunderstandings about how it should be delivered. TA support was also found to be essential for enabling class differentiation through supporting the computer games during the group work component of the sessions.

How well (actual, based on previous trials): Evidence of implementation variability.

Significant variation in the proportion of the programme content delivered was observed in the efficacy trial, although this issue appears to have been less prevalent in the effectiveness trial. Significant variation in the use of computer games was, however, observed across both the efficacy and effectiveness trials, with a fifth of pupils in the latter found to have played no games at all. Some variation in grouping practices and use of extension activities were also observed in the effectiveness trial.

Theory of Change

The ToC for the Mathematical Reasoning programme is shown in Figure 1. It outlines the target population (all Year 2 pupils) and the activities, outputs, and short-term outcomes that are intended to ultimately lead to more pupils' enhanced mathematical development and improved mathematical performance. An additional strand of outputs and outcomes reflects the intended trajectory for teacher and TA learning and broader impact.

While an initial distinction has been drawn between the whole-class and group activities to clarify that they are distinct programme components, the model articulates the manner in which the whole-class session informs group allocation, as well as their distinct but complementary contribution to the same set of outcomes. Similarly, while quantitative reasoning and number sense are treated within the programme as separate strands and developed as distinct

abilities, they are positioned as complementary skills feeding into broader mathematical outcomes.

While MR is a scripted programme, it encourages teachers to be proactive in tailoring questions and support to the children's specific needs. It also relies on teachers employing a range of teaching techniques, including scaffolding and the use of manipulatives. This means that the ToC relies on teachers having the necessary existing skills to deliver the quality of teaching required following training completion. In addition to supporting these skills, the training must also be of sufficient quality to develop the teacher and TA's understanding of what quantitative reasoning (as compared to arithmetic) and the use of inverse relations look like in practice, as the programme relies on supporting children to approach questions in a particular way, not simply to obtain the right answer.

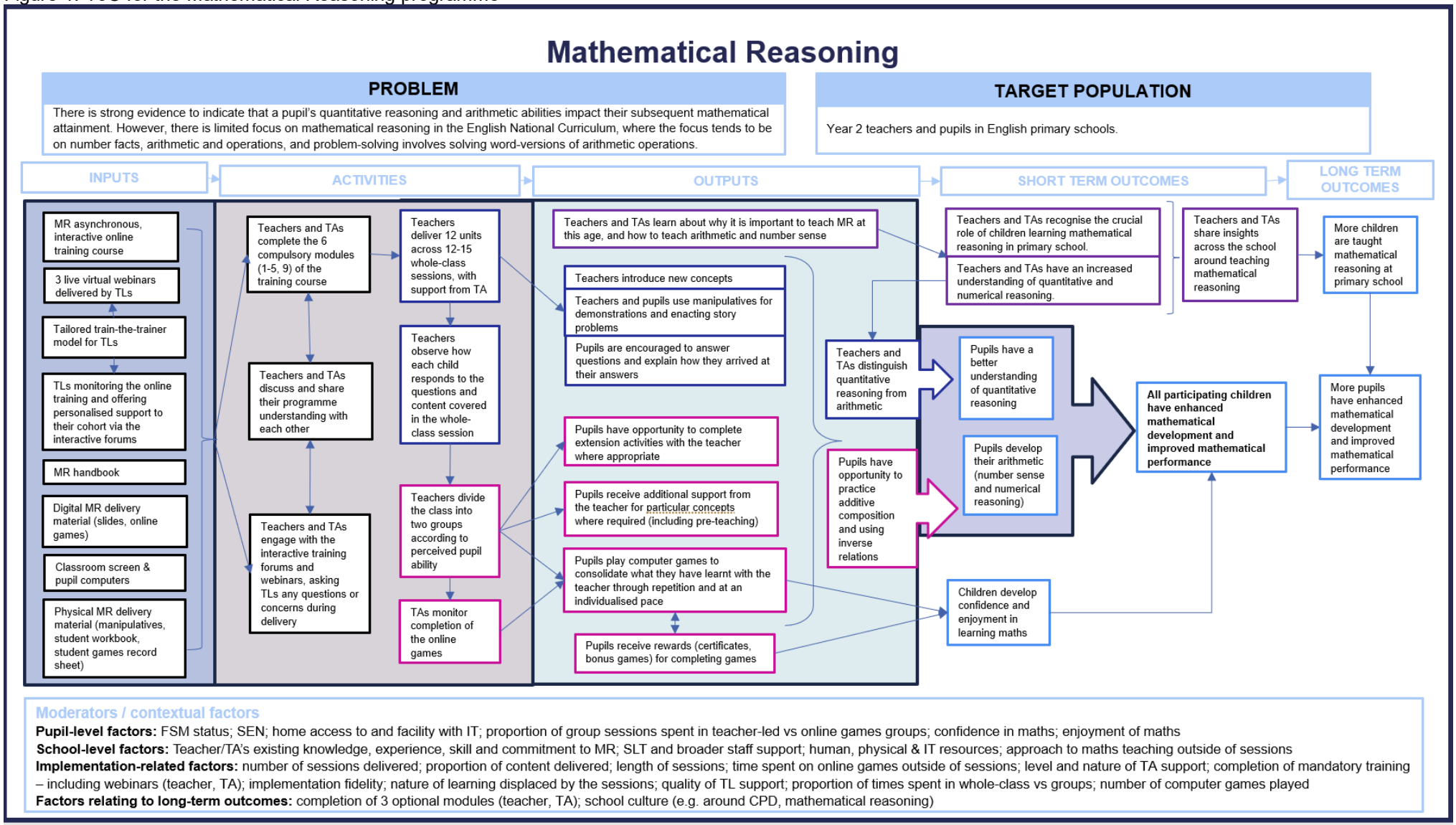
Given that the programme sessions are intended to replace normal lessons, the ToC relies on the assumption that the learning pupils receive through the programme will be more beneficial for their mathematical development than the learning they would otherwise have received through their usual lessons.

The role of differentiated teaching within the programme relies on teachers being able to correctly identify whether a pupil would most benefit from extension activities or additional support in each lesson. The effectiveness of this approach also depends on pupils in both groups drawing equal benefit from tailored teacher-led support.

Technology has a central role in the programme, which relies on schools having the necessary technology available. Moreover, the number of individual devices available would need to be sufficient so as not to influence the number of children allocated to each group. Internet access within the school would also have to be of sufficient quality to allow this number of children to access the Internet at the same time.

Finally, to achieve the programme's intended long-term outcomes, teachers and TAs must have the opportunity and means to share their learning within the school community, and other staff members must be willing and able to take this on board and integrate it into their own practice.

Figure 1: ToC for the Mathematical Reasoning programme



Impact evaluation design

Research questions

The **primary impact research question** asks:

> **RQ1**: *What is the impact of the Mathematical Reasoning programme on Year 2 pupils' attainment in mathematics (measured using GL PTM7)?*

RQ1 is aligned with the programme ToC, which identifies 'improved mathematical performance' as a short-term outcome for pupils. The question is also very similar to the primary impact research question in the previous effectiveness trial (Stokes *et al.*, 2018), including the use of the same outcome measure. Answering this will, therefore, also allow for some comparability of findings.

Secondary impact research questions focus on FSM-eligible pupils, subscales, dosage, and how impacts may vary by pupil prior attainment and computer game usage. The first secondary research question is similarly concerned with mathematical attainment measured using GL Assessment's Progress Test in Maths (PTM7) but focuses specifically on FSM-eligible pupils, as the programme ToC hypothesises FSM status to be a moderator of treatment effects.

> **RQ2.1**: *What is the impact of the Mathematical Reasoning programme on Year 2 FSM-eligible pupils' attainment in mathematics (measured using GL PTM7)?*

The second secondary research question, RQ2.2 (below), uses the same outcome measure as the primary research question but will measure impact against the PTM7 subscales. This analysis will provide additional findings for consideration alongside those of the primary research question, particularly to understand how the programme's impact is associated with specific aspects of mathematical attainment. One of the subscales in particular (mathematical reasoning) is hypothesised by the ToC to be directly associated with receipt of the programme, with the expectation that pupils participating in the programme will have a better understanding of quantitative reasoning.[8]

> **RQ2.2***: What is the impact of the Mathematical Reasoning programme on each of the PTM7 'process' categories (subscales): (i) fluency in facts and procedures, (ii) fluency in conceptual understanding, (iii) problem-solving, and (iv) mathematical reasoning (measured using GL PTM7)?*
>
> ***a)*** *for all pupils*
> ***b)*** *for FSM-eligible pupils*

[8] The developers of the MR programme specified that their definition of mathematical reasoning is not the same as that used by the PTM7 test designers.

The third and fourth secondary research questions concern dosage in order to understand how impacts vary by the number of sessions attended by the pupil and by pupils' use of computer games (which the ToC identifies as a potential moderator of impacts)[9].

> **RQ2.3:** *How does the impact of the Mathematical Reasoning programme on pupils' attainment in mathematics vary by the number of sessions attended by the pupil?*
>
> ***a)*** *for all pupils*
> ***b)*** *for FSM-eligible pupils*

> **RQ2.4:** *How does the impact of the Mathematical Reasoning programme on pupils' attainment in mathematics vary by i) the number of computer games played by the pupil and ii) the number of different computer games played by the pupil?*
>
> ***a)*** *for all pupils*
> ***b)*** *for FSM-eligible pupils*

We will also investigate whether and how the impact of the programme varies by pupil prior attainment. Although our primary research question (RQ1) includes prior attainment as a baseline to control for it and to increase precision, this question specifically looks at how programme participation interacts with pupil prior attainment. Answering it will help us to understand (along with FSM-eligible status) for whom the programme is most effective.

> **RQ2.5**: *How does the impact of the Mathematical Reasoning programme on pupils' attainment in mathematics vary by pupil prior attainment?*
>
> ***a)*** *for all pupils*
> ***b)*** *for FSM-eligible pupils*

We will investigate the impact of teacher completion of the MR training as a research question, separate to the compliance analysis.

> **RQ2.6**: *How does the impact of the Mathematical Reasoning programme on pupils' attainment in mathematics vary by whether their teacher completed the training?*
>
> ***a)*** *for all pupils*
> ***b)*** *for FSM-eligible pupils*

Design

Table 1 below gives an overview of the trial, which uses a two-arm cluster randomised design (with randomisation at the school level). The design is similar to previous trials of the same programme, thus allowing for some comparability of findings.

[9] In addition to RQ2.3 we will undertake exploratory analysis to investigate any potential relationships between when computer games are played (in school, or at home), FSM-eligibility and outcomes, as well as between computer game scores (i.e. 100% correct responses) and the pupil's outcome measure performance.

Table 1: Trial design

Trial design, including the number of arms		Two-arm, cluster randomised
Unit of randomisation		School
Stratification variables (if applicable)		N/A
Primary outcome	**Variable**	Maths attainment
	Measure (instrument, scale, source)	Progress Test in Maths 7 (PTM7) (test administered by NFER test administrators, 0-43 raw score, GL Assessment)
Secondary outcome(s)[10]	**Variable(s)**	Maths attainment (process categories/subscales)
	Measure(s) (instrument, scale, source)	Progress Test in Maths 7 (PTM7) 'process' categories (subscales): (i) fluency in facts and procedures (0-6 raw score) (ii) fluency in conceptual understanding (0-13 raw score) (iii) problem-solving (0-4 raw score) (iv) mathematical reasoning (0-20 raw score).
Baseline for primary outcome	**Variable**	Maths attainment
	Measure (instrument, scale, source)	Progress Test in Maths 6 (PTM6) (test administered by teachers, 0-31 raw score, GL Assessment)
Baseline for secondary outcome	**Variable**	Maths attainment
	Measure (instrument, scale, source)	Progress Test in Maths 6 (PTM6) 'process' categories (subscales) [11]: (i) fluency in facts and procedures (0-2 raw score) (ii) fluency in conceptual understanding (0-12 raw score) (iii) problem-solving (0-3 raw score) (iv) mathematical reasoning (0-14 raw score).

A school randomised design is consistent with the previous effectiveness trial, is straightforward to implement, and does not create any barriers to programme delivery. Randomly assigning pupils to the programme would not have been possible given the whole-class nature of the programme. Schools randomised to the control condition will continue with usual teaching (i.e. business as usual). This assumption is the specific focus of IPE RQ 5, which considers programme differentiation. All control schools will receive a £500 incentive

[10] For RQ2.2. The other secondary RQs use the same outcome as primary RQ1.

[11] For RQ2.2. The other secondary RQs use the same baseline as primary RQ1.

payment to maintain their engagement in the trial. Randomisation will not be stratified (see the 'Randomisation' section below for further details).

The primary outcome is a measure of attainment in mathematics, assessed using GL Assessment's PTM7. This is also the case for secondary research questions RQ2.1 and RQ2.3 to RQ2.6. RQ2.2, however, uses the process categories (subscales) of the same measure, resulting in four sub-outcomes. These are (i) fluency in facts and procedures, (ii) fluency in conceptual understanding, (iii) problem-solving, and (iv) mathematical reasoning. The number of sessions attended and units delivered will be measured through session delivery logs completed by the teachers.

Participant selection

All state primary schools in England are eligible to participate in this evaluation, excluding those participating in the Maths-Whizz and Maths Mastery trials or who participated in the 2023 pilot study of the new MR training model (report forthcoming). Recruitment will be nationwide and will not prioritise specific regions.

TLs will be primarily responsible for recruitment. They will approach schools via both promotional and personalised emails, events in the TLs' local areas, flyer distribution at other relevant events, and social media. The trial will also be advertised on the websites of the EEF, NFER and the University of Oxford, as well as the Reasoning First Website[12], which is maintained by the Oxford University team. Once a school has joined the project, teachers and TAs will be nominated by the school in accordance with the procedures indicated in a Memorandum of Understanding.

It is expected that most schools will choose to enter one class into the evaluation, but schools may choose to enter more than one if they wish. In addition, while most classes are expected to comprise Year 2 pupils only (the target group for the programme), mixed year group classes covering the target year group (e.g. Year 1/2, Year 2/3) are also eligible to participate. This is in order to ensure that small schools (e.g. in rural areas) have the opportunity to participate. Whilst pupils from year groups other than Year 2 may participate in the programme, only outcomes for Year 2 pupils will be assessed.

Pupils will complete a baseline assessment to control for prior attainment and increase the precision of the analysis, but baseline assessment scores will not be used to determine pupil eligibility for the trial. We expect approximately 3,084 pupils to receive the programme in total.

Outcome measures

Primary and secondary outcomes

The **primary outcome** for the trial is maths attainment, measured using overall raw scores from GL Assessment's PTM7 (Form A). For this trial, we expect raw scores to be less

[12] The MR programme website (https://reasoningfirst.org.uk/)

vulnerable to floor and ceiling effects than age-standardised scores[13]. The choice of raw score rather than age-standardised score also makes the primary and secondary outcomes more comparable (age-standardised scores are not available for the PTM7 subscales). There will be no adjustment for pupil age, either through standardisation or by including age as a model covariate, given that the maximum age gap between any two trial pupils is roughly one year. Randomisation will help ensure that results are not biased by age imbalances between the trial arms, and we do not expect age to explain a large amount of model variance after accounting for the PTM6 baseline.

The choice of 'maths attainment' for the primary outcome is consistent with the ToC which includes 'improved mathematical performance' as a short-term pupil outcome. GL Assessment's PTM offers a suitable measure for the primary outcome of mathematical performance, particularly as it aligns with national curriculum objectives in English schools. Furthermore, the use of PTM7 is consistent with the previous effectiveness trial, thus allowing for a degree of comparability with previous evaluations of the programme.[14]

Multiple versions of the PTM assessment are available, representing different levels (i.e. year groups). This means it can be used to measure progress within a single year or across multiple years, thus making it suitable for both a baseline and outcome measure in this study. The PTM questions were developed by the Mathematics Assessment Resource Service (MARS) team, which is a collaboration between the University of California, Berkeley and the Shell Centre, Nottingham. Standardisation was conducted with 34,762 pupils in the UK (for all test versions; 4,071 pupils for PTM7). Internal consistency is reported at 0.91 (Cronbach's Alpha), and the assessment has been found to have no significant difference in standard age scores between male and female pupils. Given the robust development process of the test and its established psychometric properties, we will use the instrument as developed by GL Assessment and without modification for our main analysis (see also 'Additional analyses and robustness checks' section below).

All PTM assessments are based on categories of mathematical proficiency, which have been derived by GL Assessment from the Curriculum Aims in the KS1, KS2 and KS3 National Curriculum for England (2013). They are also comparable with the GCSE Assessment Objectives. The assessment comprises of 36 questions aligned with these categories, as follows (further details in Appendix B):

- Fluency in facts and procedures
- Fluency in conceptual understanding
- Mathematical reasoning
- Problem-solving

[13] This is based on inspection of the distribution of baseline PTM6 raw and age-standardised scores for this trial, as well as comparing figures 3 and G.1 in the previous effectiveness trial.

[14] The efficacy trial used Progress in Maths 7 (PiM7), the forerunner to PtM7. GL Assessment report the correlation between the PiM7 and PTM7 to be 0.8, based on a sample of 350 pupils.

The categories will be used as subscales for **secondary outcomes**,[15] thus providing additional findings for consideration alongside those of the primary question, particularly to understand how the programme's impact is associated with specific aspects of mathematical attainment. One of the subscales in particular (mathematical reasoning) is hypothesised to be directly associated with receipt of the programme and is identified by the ToC as a short-term outcome ('Pupils have a better understanding of quantitative reasoning'). However, the developers of MR emphasise that the programme promotes reasoning about relations between quantities and about relations between numbers. This is a narrower focus than the 'mathematical reasoning' PTM7 subscale; not all questions on the subscale relate to programme content.

The PTM7 subscales - as defined by GL Assessment - will be used in the secondary analysis, rather than developed using data-driven methods such as exploratory factor analysis. We will analyse the trial test data for reliability as a sensitivity check (see confirmatory factor analysis in the analysis section) and expect it to be sufficient given GL Assessment's prior development work and published data.

Assessments will be administered at the endpoint by NFER Test Administrators, who will be blinded to school allocation to programme or control.[16] Pupils will complete the assessment within the classroom and under test conditions. PTM7 is not time-limited, but GL Assessment has suggested that approximately 35 minutes would be needed for pupils to demonstrate their abilities. Test administrators will use a secure courier to send the assessment papers to GL Assessment, who will complete the marking and scoring, blinded to group allocation. GL Assessment will then use the NFER secure portal to share with NFER a spreadsheet containing the pupil-level data, including overall raw scores and sub-total raw scores for each of the aforementioned categories, alongside item-level scoring. Standard Age Scores, Stanine Scores and National Percentile Ranks will also be included in this data.

Baseline measures

The baseline measure will be GL Assessment's PTM6 (for RQs 1, 2.1, 2.3, 2.4), which has been designed for administration at the start of Year 2. This assessment is based on the same categories as PTM7, with a Pearson correlation between PTM6 and PTM7 of 0.67 (Bishenden, 2023). For RQ2.2, the baseline measure will be the PTM6 subscales.

The baseline assessment will be administered by class teachers, who will be asked to do so within the classroom and under exam conditions. NFER will coordinate for a secure courier to deliver the assessments to GL Assessment to complete the marking and scoring. GL Assessment will then share the pupil-level data with NFER via NFER's secure portal.

[15] For RQ2.2. The other secondary RQs use the same outcome as primary RQ1.

[16] While schools will be asked not to share this information with the Test Administrator, it is possible that the Test Administrator may be made aware of the school's allocation through interaction with pupils and/or staff.

Sample size

Table 2 below illustrates our estimation of the minimum detectable effect size (MDES) given a randomised sample of 240 schools, with a 1:1 allocation of schools to the two trial arms. The headline MDES estimates assume both school and pupil-level attrition (detailed below, based on our experience of running similar trials), but for illustration and comparison, we have also provided further MDES estimates that assume no attrition. Given that the previous MR effectiveness trial (Stokes *et al.*, 2018) found an effect size of 0.08 and 0.09 for all pupils and FSM-eligible pupils, respectively. We think it is appropriate to power this trial for a relatively low MDES. We have therefore recommended that the Oxford University team recruit schools to the upper limit of their capacity for delivery (120 schools).

Table 2: Sample size calculations (240 schools recruited)

		Overall	FSM-eligible
Minimum Detectable Effect Size (MDES)	**Assuming attrition**	**0.117**	**0.154**
	Assuming no attrition	0.108	0.140
Pre-test/ post-test correlations	level 1 (pupil)	0.62	0.62
	level 2 (school)[17]	0.62	0.62
Intracluster correlations (ICCs)	level 2 (school)	0.11	0.11
Alpha		0.05	0.05
Power		0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided
Average cluster size		25.7	6.8
Number of schools	Intervention	120	120
	Control	120	120
	Total	240	240
Number of pupils	Intervention	3,084	816
	Control	3,084	816
	Total	6,168	1,632

[17] This assumes that the proportion of school-level variance explained by the pre-test (that is, the proportion reduction in level 2 variance if the pre-test is removed from a model) will be approximately the same as the proportion of pupil-level variance explained. We note that the figures in appendix table A1 of the efficacy trial support this assumption.

The table assumes one class per school (25.7 pupils per class, 7 of whom are FSM-eligible[18]), with 10% of schools having mixed-year classes (which would mean fewer pupils eligible for the evaluation in these schools). While a small number of schools may wish to enter more than one class (a mixed year or otherwise), we have not included this assumption in our estimates as it is likely that the occurrence of this will be low given the staffing constraints schools are currently facing. We, therefore, expect that the sample size may be larger than our estimate, resulting in a higher level of statistical power for the study.[19] We have assumed 10% school-level attrition and 15% pupil-level attrition within those schools that remain due, for example, to absences on the day of assessment.

While GL Assessment reports that the correlation between PTM6 and PTM7 is 0.67, we recommend basing estimates on a slightly lower correlation (0.62, as illustrated). This is due to the use of these assessments in the context of a programme, where we would expect different correlations between the intervention and control groups, resulting in an overall lower correlation than what would be found outside of a programme context. We have also assumed a school-level ICC of 0.11, which aligns with analysis from the previous effectiveness trial (Stokes *et al.*, 2018) and with the findings of Singh *et al.* (2023) based on an NPD sample and a range of EEF studies.

The MDES values given in Table 2 were calculated using the PowerUpR package in the R statistical software, using the function 'mdes.cra2'.

Randomisation

We will simultaneously randomise the recruited schools on a 1:1 basis. The randomisation will be done using R statistical software and syntax developed specifically for this evaluation by an NFER Statistician. This will allow for transparency and replicability of the randomisation process and will facilitate the recording of allocations.

We will perform simple randomisation without stratification. Simple randomisation is considered to be safer than restricted (i.e. stratified) randomisation, which may increase the chance of selection bias and errors during the randomisation process (Hewitt and Torgerson, 2006). Stratification may also be difficult to implement if there are multiple stratifiers and where cells produced by stratification include no units. Whilst not stratifying, we have specified a robustness check, which will include pupil-level characteristics that may be important predictors of the outcome (e.g. SEND status, FSM eligibility).

To assess balance at baseline, we will produce cross-tabulations of pupil and school characteristics (e.g. number of classes) in the sample. The pupil-level characteristics will be:

- FSM eligibility

[18] 25.7 as we assume that 10% of schools will have mixed year classes. 26.9% of pupils indicated as eligible for FSM by the EVERFSM_6_P NPD variable. 2022/23 DfE figures indicate 24.6% of Year 2 pupils were FSM eligible, a figure that has risen year-on-year.

[19] Note that a maximum of 240 schools will be recruited to the trial as this is the upper limit of what the delivery model currently has the capacity to support.

- SEND status
- Gender
- EAL status
- PTM6 raw score (comparison of means)

We expect that this check will confirm the correct functioning of the randomisation, but we will include a robustness check which replicates the model used to estimate RQ1, while also controlling for the above pupil-level characteristics

Statistical analysis

Primary analysis

The primary outcome of PTM7 score will be the dependent variable in a multilevel linear regression, with two levels (pupil, school). An indicator variable for group allocation (intervention or control) will be used to measure the causal impact of Mathematical Reasoning, with baseline PTM6 score included as a covariate (RQ1). This analysis will be 'intention-to-treat', with pupils analysed according to their randomisation status, regardless of their level of participation in the MR programme. Analysts will not be blind to group allocation in this or any other part of the impact analysis.

Secondary analysis

As outlined in the outcomes section above, four subscales of the PTM7 will be included as secondary outcomes (for RQ2.2 only – other secondary research questions will use the primary outcome of PTM7): (i) fluency in facts and procedures, (ii) fluency in conceptual understanding, (iii) problem-solving, and (iv) mathematical reasoning. The four PTM7 subscales will be the dependent variables in four linear two-level (pupil, school) regressions. As in the primary analysis, the impact of the programme on each outcome will be estimated using an indicator variable, and the appropriate baseline PTM6 subscale score will be included as a covariate.

As there are multiple outcomes in the secondary analysis, this will lead to an inflated 'family-wise error rate' (chance of one or more false positives) when conducting significance tests. We do not, however, intend to implement a multiple testing correction (e.g. Bonferroni) to address this. This is because it is likely that the interpretation of 'corrected' p-values would be extremely underpowered, given the sample size restrictions of this evaluation. We will be following the EEF guidelines on reporting statistical significance (focus on point estimates and confidence intervals, do not interpret p-values in a binary way), which renders multiple testing corrections less important.

Sub-group and moderator analyses

The effect of the programme amongst FSM-eligible pupils will be determined by repeating the primary analysis model for this subgroup (RQ2.1). The NPD variable 'EVERFSM_6_P', which indicates whether a pupil has been eligible for FSM in the last six years, will be used to identify FSM pupils throughout the impact analysis. Additionally, we will investigate the differential

effect of the programme for FSM-eligible pupils relative to non-FSM pupils by repeating the primary analysis model with an indicator for FSM eligibility added, as well as an interaction term between FSM and the programme indicator. This interaction term represents the differential effect of the programme amongst FSM-eligible pupils.

For RQ2.2b, RQ2.3b, RQ2.4b and RQ2.6b, we will repeat the analytical approach used in part (a) (i.e. for all pupils) of those questions but using a restricted subgroup of FSM-eligible pupils. Further analysis will be performed investigating how the impact of the intervention varies with prior attainment (RQ2.5a), which will be included as a continuous variable. This will be done by adding an interaction term between the baseline PTM6 score and group indicator to the primary analysis model, which represents the differential impact of the intervention as prior attainment increases. This same analysis will also be repeated for only FSM-eligible pupils, investigating the differential impact of the intervention as prior attainment increases for this subgroup. This second analysis, restricted to FSM pupils, is exploratory as it is very likely to be underpowered, so results will be caveated in the final report.

The compliance analysis described below focuses on pupil attendance of each programme unit. Separate analyses will investigate the impact of the number of sessions pupils attend (RQ2.3) and whether their teacher completes the MN training (RQ2.6) on PTM7 scores. The analysis for RQ2.3 and RQ2.6 will use two-level (pupil, school) linear models, similarly to the primary and secondary analysis, rather than taking an instrumental variable approach.

Compliance and dosage analysis

Teachers at intervention schools will complete a delivery log of MR sessions. For each session, the unit completed in that session will be recorded (or 'unit not complete' if no unit was completed), together with which pupils attended the session. By combining this information, it is possible to count the number of units attended by each pupil and across how many sessions. Where a unit spans multiple sessions, a pupil's attendance at all those sessions will be required for the unit to be counted. A histogram showing the number of MR units attended by pupils will be included in the final report. There will be three compliance measures for this evaluation, all of which are based on a pupil's attendance of the 12 units delivered by their teacher.

1. Number of units attended by a pupil (continuous measure)
2. Whether a pupil attended 10 or more units (dichotomised variable derived from continuous measure)[20]
3. Whether a pupil attended one or more units (dichotomised variable derived from continuous measure)

[20] By default this will be compliance measure (2). However, if fewer than 40% of intervention pupils meet this condition we will instead use the median number of units as the compliance cut-off for measure (2). This is to insure against a scenario where very few pupils meet the compliance condition, which worsens the potential bias from exclusion restriction violations and limits the applicability of results.

For this evaluation, the main compliance measure (1) will be continuous. Compliance measure (2) investigates the impact of attending most units. Compliance measure (3) exists as a lower bound for measure (2), as described below.

The estimated impact will be obtained for the above compliance measures using instrumental variable modelling. For instrumental variable analysis to produce unbiased results the 'exclusion restriction' must hold: intervention assignment can only affect PTM7 scores via the chosen compliance measure. One implication of the exclusion restriction is that when dichotomising a continuous measure (here number of units), intervention pupils receive no benefit below the chosen compliance threshold. This is unlikely to hold for compliance measure (2): we expect some benefit for pupils attending fewer than 10 units, which is likely to upwardly bias this result. For this reason we included compliance measure (3) to act as a 'lower bound' for measure (2). Two models will be created for each compliance measure described above, one for all pupils and one restricted to FSM-eligible pupils (for a total of six models). Details will be provided in the statistical analysis plan (SAP).

Additional analyses and robustness checks

We will undertake a Confirmatory Factor Analysis (CFA) to determine whether additional sensitivity analysis is required: the PTM subscales, which are the outcome for RQ2.2, were identified by GL Assessment for general usage by mapping assessment questions to national curriculum areas. Running a CFA on the four PTM7 subscales (using a 4-factor solution, one for each subscale, with the items from each subscale loading onto one factor only) will allow us to understand factor loadings and overall model fit. Should we identify items that perform poorly for one or more subscales, we will rerun any secondary analysis model that uses an affected subscale, with the poorly performing items removed from the subscale.

As described in the 'Randomisation' section above, we will rerun the primary analysis with additional pupil-level covariates (SEND status, FSM eligibility, gender) to assess the sensitivity of the primary analysis result to chance imbalances in these variables. If these variables explain a substantial proportion of model variance, there may also be moderate improvements in the precision of the estimate. All three variables will be added to this model, regardless of the degree of imbalance between the control and intervention groups observed at baseline.

Missing data analysis

All impact analysis described above will be a 'complete case' analysis; pupils with missing outcome or predictor data for a given model will not be included in the model. The degree of missing data for each primary analysis variable will be reported; in practice, it will be baseline and endpoint PTM7 scores that are missing. Further exploration of the pattern of missing data and implications for the reliability of the primary analysis result will be conducted in accordance with the EEF analysis guidance. As part of this exploration we will perform a sensitivity analysis for a scenario where the data is 'missing not at random', as discussed in the SAP.

Where schools drop out (do not provide endpoint PTM7 data), we will try to establish the reason for this and, where possible, include a brief summary in the report.

Estimation of effect sizes

The beta coefficient for each binary predictor (conditional on model covariates) will be converted into an effect size by dividing by the square root of the pupil-level plus the school-level variance. A confidence interval for the effect size will be calculated by dividing the endpoints of a 95% confidence interval for beta by the square root of the pupil-level plus the school-level variance. Continuous predictors such as dosage will be presented on their raw scale.

Implementation and process evaluation (IPE) design

The IPE has been designed to complement the impact evaluation and to enable a stronger understanding of the mechanisms within the ToC. The mode of the training has changed since the previous trial of the MR programme based on the hypothesis that this would increase fidelity by making the training more accessible for teachers in terms of both when and where they can complete it. As a result, the IPE is particularly focused on understanding variation and adaptation in implementation, particularly in terms of the use of the online games and adherence to the programme structure – as the implementation of both of these aspects varied highly in the previous effectiveness trial. The role of differentiated teaching within the programme is also of interest, including understanding how well teachers were able to adopt the flexible and adaptive grouping approaches intended by the programme. Moderators and contextual factors will be interrogated, including access to IT facilities and the level and nature of the support from the TA and Senior Leadership Team. This will help contextualise the IE findings by providing insight into what does and does not work in relation to the programme and external factors that may influence this. Perceived impacts and possible mediating pathways will likewise be explored to develop the ToC further. Finally, the intention of the programme to both replace and complement standard maths lessons requires interrogation of the extent of both programme differentiation and the potential risk of learning displacement.

This IPE approach was confirmed as part of the set-up phase following the IDEA workshop, which offered the evaluation team the opportunity to speak in depth with the Oxford University team about the programme and the key areas of interest for the IPE.

Research questions

Fidelity, adaptation & reach

IPE RQ 1: *To what extent was the MR programme delivered as intended at scale?*

- To what extent did the training model sufficiently prepare teachers and TAs for programme delivery?
- What was the nature and extent of ongoing support for teachers and TAs (from TLs) over the course of the programme?
- To what extent did teachers and TAs engage with the support available, and what motivated this?

- To what extent was the programme delivered in accordance with the training and guidance?
- What was the nature and extent of any adaptation?
- Did the programme reach all children in participating classes?

This research question will provide key information about programme compliance for the impact evaluation and offer insights into the extent to which the training and programme were implemented with fidelity (i.e., as intended). We are particularly interested in understanding how effective the new training model is in preparing schools for programme delivery and what impact this has on the levels of fidelity observed. Practices around differentiated provision will likewise be explored to understand the extent to which this happens as intended and any implications this may have for the risk of differentiated pupil outcomes. Finally, we will explore adaptations across schools to understand the kind of variation that is introduced by delivery in a 'real-world' context and assess the extent to which this is perceived to enhance or reduce programme impact, including around the use of computer games.

Context, moderators, dosage

> **IPE RQ 2:** *What are the key moderators and contextual factors that influenced how effectively the programme was delivered?*

- What were the key challenges and facilitators to successfully implementing the programme?
- To what extent were changes made to the programme in response to the findings of previous trials effective in increasing programme fidelity? This includes new training materials and guidance.
- What was the perceived value of specific programme elements (e.g. online training, TL support, allocation of responsibilities between the teacher and TA, use of manipulatives, computer games)?
- In what way(s) and to what extent does the programme fit within schools' maths curriculum?
- What (if any) challenges or facilitators were observed that would be relevant to further scale-up of the programme in future?

This research question looks to understand any contextual or individual factors that may influence the extent to which pupils are able to benefit from the programme, including access to IT facilities and the level and nature of support from the TL. This will help to contextualise the findings of the impact analysis in relation to any barriers or facilitators that may have influenced the outcomes observed. This question will also help us to understand if there are any particularly significant moderators that need to be better accounted for in the ToC and inform future programme design and implementation guidance to ensure as many potential barriers to impact are removed as possible. Understanding the perceived relative value and/or significance of different programme components will likewise help us to further refine

the ToC. In addition, we are particularly interested in understanding how the programme relates to the schools' curricula to address the potential risk for learning displacement as a result of the loss of the normal lessons replaced by the sessions.

> **IPE RQ 3:** *How did pupils' experiences of the programme vary depending on pupil characteristics?*

- To what extent did the programme meet the needs of different pupil sub-groups (e.g. disadvantaged pupils (eligible for FSM), pupils with different levels of prior attainment, pupils with English as an additional language (EAL), pupils with SEND)?
- What challenges and/or facilitators emerged specific to one or more of the above subgroups?

This research question will provide insight into the potential reasons behind any differences that the impact analysis may find in terms of outcomes for pupils eligible for FSM and/or pupils with different levels of prior attainment. While no differential impact was seen by ability level in either of the previous trials, the IPE for the first effectiveness trial reported that teachers felt the programme was less suitable for the highest and lowest-ability learners, and there was some indication that pupils of lower prior attainment were more likely to drop out of the analysis sample (Worth *et al.*, 2015); Stokes *et al.*, 2018).

This question will also provide some insight into any perceived differential impacts for groups that the impact analysis will not be able to measure – primarily pupils with EAL and/or SEND.

Perceived impact

> **IPE RQ 4:** *What was the perceived impact of the MR programme for (i) staff members, (ii) pupils and (iii) disadvantaged pupils specifically?*

- To what extent was the programme perceived to increase teacher and TA understanding of mathematical reasoning and recognition of the importance of teaching it to this age group?
- To what extent was the programme perceived to increase pupils' mathematical reasoning, confidence and/or enjoyment of maths (including for disadvantaged pupils specifically)?
- To what extent is there potential for learning from the programme to be integrated into the long-term practice of participating staff members and pupils?
- To what extent is there potential for learning from the programme to be integrated across participating schools beyond the participating classes?
- What (if any) negative unintended consequences for pupils, staff and/or schools stemmed from programme implementation, including in relation to learning displacement?

Providing CPD around the importance of and skills for teaching mathematical reasoning is an important objective of the programme as a mediator of pupil outcomes within and beyond the programme. This research question will help us to understand the extent to which this plays out in practice. As there is no existing evidence or validated measure from the previous MR

trials relating to the teacher outcomes, this analysis will remain exploratory and within the IPE rather than a pre-specified mediator analysis of impact. We will also look to understand the extent to which the causal chain for a long-term school-level impact (see ToC) may be seen to hold true.

In addition, this question will explore whether there are any indications of an impact on child enjoyment or confidence in maths, which could constitute a supporting pathway to the intended outcomes, as proposed in the ToC. While these pupil outcomes were not assessed in the previous trials, there were reports in both cases of children enjoying the activities and demonstrating greater confidence in maths-related class participation (Worth *et al.*, 2015; Stokes *et al.*, 2018). This question will also provide further insight into any perceived differential impact for FSM-eligible pupils, complementing the impact analysis and IPE RQ3 (see above).

Finally, we will look to uncover any potential negative unintended consequences of programme implementation, such as learning displacement for participating pupils (or, indeed, staff, if they forfeit time for alternative maths-related CPD as a result), increased workload for teachers, and reduced access to IT resources and/or TA time for pupils not in the programme classes.

Programme differentiation

> **IPE RQ 5:** *What was business as usual (BAU) in relation to maths teaching, and to what extent did it differ from the MR programme?*

- What was the nature of BAU in all schools prior to the programme?
- What was the nature of BAU in control schools during the programme?
- To what extent did the programme provide different opportunities for participating pupils compared to BAU?

This question will complement the impact analysis by looking to understand the extent to which practice in control and treatment schools differed and, hence, the extent to which any difference in outcomes may be attributable to the programme. It will also inform our understanding of which components of the ToC represent a genuine departure from what schools already have in place. Finally, this question will inform our understanding of the extent to which the programme is both differentiated from standard maths lessons and aligned with the curriculum, as is the programme's ambition, particularly given that some teachers in the previous efficacy trial expressed concerns about a dissonance with encouraged school practice (Stokes *et al.*, 2018).

Research methods

We are planning to use a varied mixed-method approach to capture the IPE data, as described below (see also Table 3 for a summary of the IPE methods).

The IPE will follow a multi-phase design, with interleaved qualitative and quantitative elements and emerging findings informing the design of subsequent data collection instruments. The sequencing of the IPE activities is outlined in Figure 2 below. The pre-trial BAU survey will help us to understand the extent to which the programme differs from standard practice in

maths lessons, while the post-delivery BAU survey allows us to determine the extent to which practice in control schools differed from practice in intervention ones. As the BAU and Teacher & TA surveys both include measures of teacher and TA outcomes, this sequencing allows for pre-post testing and comparisons between intervention and control. Conducting the structured observations and reviews of the TL and teacher/TA training ahead of the case studies means findings from the former can feed into the development of data collection tools for the latter. Similarly, findings from the case studies and data collection around training and support will inform the development of the Teacher & TA survey, which will seek to obtain an overview of the issues and experiences of schools in implementing the programme. The TL interviews will occur at intervals during the delivery period to enable a range of perspectives to be gathered while also being able to focus on different stages of implementation.

Figure 2: Sequencing of IPE data collection activities

Pre-trial	During training	During delivery	Post-delivery
BAU survey (all)			BAU survey (control)
Semi-structured observation of TL training	Structured review of online training course	Semi-structured webinar observations & TL interviews	Training course completion & webinar attendance data
		Case study observations, interviews & focus groups *(t1)*	Case study interviews *(t2)*
		Session delivery & attendance data Computer games data	Teacher & TA survey (intervention)

A summary of the IPE methods and how they relate to the IPE dimensions and research questions can be found in Table 4.

Table 3: IPE methods overview

IPE dimension	RQ addressed	Research methods	Data collection methods	Sample size and sampling criteria	Data analysis methods
Fidelity	1	Review	Semi-structured review	Teacher & TA online training content	Qualitative content analysis
		Observation	Semi-structured observations	1x TL training sessions; 6 x webinars; 8 x programme sessions (across 8 treatment schools)	
		Interviews	Semi-structured interviews	6 x TLs; 8 x teachers, 8 x TAs, 8 x maths leads (across 8 treatment schools)	
		Focus groups	Focus groups	8 focus groups with 4 pupils (across 8 treatment schools)	
		Surveys	Online questionnaire	Teacher & TA endpoint survey (treatment teachers & TAs)	Descriptive statistics
		Monitoring information	Training course completion data	All treatment schools	
			Webinar attendance data	All treatment schools	
			Session delivery & attendance logs	All treatment schools	
			Computer games data	All treatment schools	
Context & moderators	2, 3	Review	Semi-structured review	Teacher & TA online training course resources	Qualitative content analysis
		Observation	Semi-structured observations	6 x webinars; 8 x programme sessions (across 8 schools)	

		Interviews	Semi-structured interviews	6 x TLs; 8 x teachers, 8 x TAs, 8 x maths leads (across 8 treatment schools)	
		Focus groups	Focus groups	8 focus groups with 4 pupils (across 8 treatment schools)	
		Surveys	Online questionnaires	Teacher & TA endpoint survey (treatment teachers & TAs)	Descriptive statistics
		Monitoring information	Training course completion data	All treatment schools	
			Webinar attendance data	All treatment schools	
			Session delivery & attendance logs	All treatment schools	
			Computer games data	All treatment schools	
Perceived impact	4	Interviews	Semi-structured interviews	8 x teachers, 8 x TAs, 8 x maths leads (across 8 treatment schools)	Qualitative content analysis
		Focus groups	Focus groups	8 focus groups with 4 pupils (across 8 treatment schools)	
		Surveys	Online questionnaires	Baseline BAU survey (all teachers); Endpoint BAU survey (control teachers); Teacher & TA endpoint survey (treatment teachers & TAs)	Descriptive statistics; prep-post control vs programme
Programme differentiation	5	Observation	Semi-structured observations	8 x programme sessions (across 8 schools)	Qualitative content analysis
		Interviews	Semi-structured interviews	6 x TLs; 8 x teachers, 8 x TAs, 8 x maths leads (across 8 treatment schools)	

		Surveys	Online questionnaires	Baseline BAU survey (all teachers); Endpoint BAU survey (control teachers); Teacher & TA endpoint survey (treatment teachers & TAs)	Descriptive statistics
		Monitoring information	Session delivery & attendance logs	All treatment schools	
			Computer games data	Computer games data	

1. Reviews and observations of training and support material

The nature and quality of training and support provided to schools will be explored through the triangulation of four different data collection processes spanning both levels of the blended learning approach (TL training and teacher/TA training). The data collection instruments for these activities will be informed by the programme ToC, TIDieR framework and the IPE research questions, with a particular focus on fidelity. These data collection activities will inform our understanding of the nature and accessibility of support available, as well as the frequency and nature of engagement with it.

Semi-structured observation of TL training

An NFER researcher will attend and observe the second in-person TL training day, which focuses on preparing the TLs to support programme implementation. The observation will be semi-structured, with the researcher noting down points of interest in relation to the training structure and content based on a pre-determined topic guide developed from the programme material. The data from this observation will provide important context for later data collection activities around fidelity, moderators, and the nature and quality of support teachers and TAs receive.

Review of the teacher & TA online training course resources|

An NFER researcher will complete a structured review of the online training course material. This review will be completed in parallel to training completion by the schools and will inform development of the instruments for subsequent data collection activities – particularly for the webinar and session observations.

Semi-structured interviews with TLs

Remote interviews via video call will be carried out with six of the TLs. Each interview will last approximately 45 minutes. The timeline for these interviews will be approximately mapped onto the webinar delivery schedule, such that two TLs will be interviewed following each set of webinars. One more experienced TL and one less experienced TL will be interviewed in each instance. These interviews will seek to gather further detail about the TL training and preparation process, as well as to gauge TL's understanding of their role and the nature of the support they are providing. We are particularly interested in understanding how well-equipped the TLs feel themselves to be to support the implementation of the programme, as well as what this support looks like in practice. In addition, we will look to understand the kinds of questions and requests schools are making to understand better the areas where schools may be facing more challenges, as well as the extent and nature of school engagement with both the programme and the support available.

Semi-structured observations of webinars

An NFER researcher will attend and observe the webinar of two different TLs for each of the three webinar types, ensuring that both a more and a less experienced TL is covered each time. Particular attention will be paid to how the webinars complement the online training material in terms of both additional support and consistent messaging. We will also look at the nature of participant engagement, including the nature of questions asked during the webinars, as well as more qualitative aspects such as the level and quality of interaction and perceived

levels of satisfaction and/or concern with the programme. As these webinars will be spaced out at intervals across the delivery period, we will note any observations relating to the evolution of these elements over the course of the programme, as well as how they compare between different cohorts. We will also be able to compare how different TLs approach their webinar and their cohort more broadly, and any differential dynamics this may create. In addition, questions raised by attending practitioners will inform our analysis of context and moderators that informed implementation, particularly in terms of challenges and facilitators that teachers and TAs observed. As with the TL training, these observations will be semi-structured, with the researcher noting down points of interest in relation to a pre-determined topic guide.

2. Case studies

Case studies will be carried out with eight schools over the course of the delivery period to help us develop an in-depth understanding of numerous IPE dimensions. This number will allow us to achieve sufficient variety in size, location, allocated cohort and TL experience, while minimising the number of schools participating in additional activities. We will prioritise schools with an overrepresentation of FSM-eligible pupils for case study selection to optimise the opportunity to explore how the programme affects this group of pupils. We have chosen not to sample based on engagement level with the aim of understanding a range of experiences, including barriers to engagement. Case study schools will receive a 'thank you' payment of £100 in recognition of the additional time they have given to the research, as well as to encourage participation from schools regardless of their level of engagement in the programme.

Specific schools will be invited to participate in the case studies on the basis of the above criteria (although they may choose to decline). We will clearly communicate what being a case study schools involves and what is asked of which staff members before confirming the school as a case study. We will work closely with the main contact at the school to agree a timetable for the case study activities.

Data collection will take place at two-time points for each case study school. *T1* data collection will be primarily in-person and take place approximately midway through the delivery period (February-March 2025) and will focus on the experience of the training and programme implementation. *T2* will take place remotely in parallel to endpoint data collection (June 2025) and will focus on perceived outcomes, including any indications of broader CPD-related impact, as well as any challenges, facilitators and/or adaptations that emerged over the remainder of the implementation period. Data collection from multiple sources in each case study (teachers, TAs, maths leads, and pupils) will also allow us to triangulate perspectives to understand the nuances of implementation better.

Structured observation of whole class and group session delivery (t1)

An NFER researcher will visit each case study school at *t1* to observe the delivery of one session, including both the whole-class and group components. During the group component, we will focus on the teacher-led group but still look to assess the level of pupil engagement with the computer games and the nature and extent of TA support.

The aim will be to observe sessions at different delivery stages to understand the breadth of experience. This will also allow for flexibility to reduce the burden on participating schools. Regarding their place within the programme, all sessions are compulsory and comparable in their purpose and structure.

The researcher will use a structured observation tool adapted from the Oxford University team's Fidelity Scale.[21] The tool will look at fidelity and adaptations to the teaching techniques outlined in the training course, adherence to the stipulated session length and structure, contextual considerations such as equipment and location, and qualitative assessments of pupil engagement.[22] We will also observe the process by which pupils are divided into the two groups and the proportion of children in each. The data collected with the tool will be qualitative in nature.

Teacher and TA interviews (t1 + t2)

Teacher and TA interviews will occur at both time points (*t1* and *t2*). Interviews at *t1* will occur in person on the day of the observation visit and last between 30 and 45 minutes. We will ask about participant experiences of the programme training and support, how they have found delivering the programme, and any challenges, facilitators, or relevant contextual factors they have encountered. In addition, teachers will be asked about their pupil grouping practices and any adaptations they have made for programme delivery, along with the rationale.

The interviews at *t2* will take place remotely via video call and last approximately 30 minutes. These interviews will focus on perceived outcomes (whether intended or otherwise) for the pupils and themselves, as well as any indications of broader, long-term outcomes for the school due to knowledge-sharing and raised awareness of the importance of teaching mathematical reasoning from a young age. We will also probe about any potential negative unintended consequences that may have emerged, such as learning displacement. The interviews will also cover any other programmes and interventions delivered to Year 2, to supplement the BaU survey (below).

Pupil focus groups (t1)

As part of the *t1* visit, we will ask teachers (in advance) to nominate four children to participate in a brief focus group following the session, with the request that they include some pupils eligible for FSM and covering both groups L1 and L2 of the programme. The NFER researcher will work with these children in an appropriate space selected by the teacher and in the presence of another member of the school staff. The focus group will last for up to 15 minutes and will use age-appropriate visuals and creative methods to explore the children's views of the MR session and their feelings about maths. For each programme component (whole-class

[21] The Fidelity Scale was piloted as part of a Masters project and was found to have a high level of inter-rater reliability (kw = 0.811, $p < 0.001$) (Yao, unpublished).

[22] This refers to the extent to which the pupil appears to be listening (not distracted or talking on unrelated topics) and actively contributing (responding to questions, raising their hand, etc.).

component, teacher-led group component and computer games), focus group pupils will be asked to circle one face (happy, neutral, sad) for how they felt about it. They will then be asked follow-up questions about what they liked most and least about each component and whether they felt they were good at it. Finally, we will probe how these feelings compare to how they feel in other maths lessons.

Maths lead interviews (t1 + t2)

The maths lead in each case study school will be invited to participate in a brief interview at both time points. As with the teacher and TA interviews, the interview at *t1* will be ideally conducted in person on the day of the visit (we will offer remote interviews if this is not possible). This interview will ask about the school's motivation for engaging with the trial, as well as broader challenges and facilitators that have been encountered in implementing it. A follow-up remote video call interview will be conducted at *t2* to explore the perceived outcomes of the programme for the broader staff body and indicators of any longer-term impact in accordance with the programme ToC.

Where the maths lead is also the teacher delivering the programme, relevant additional questions will be added to the teacher interview; no separate maths lead interview will be conducted.

3. Online surveys

BAU survey

Two online BAU surveys will be implemented: one at baseline (teachers in all schools) to inform our understanding of BAU in all participating classes prior to programme delivery, and one at endpoint (teachers in the control group only) to understand what BAU looked like in control schools over the trial period. The BAU survey will ask about usual practice in maths lessons, maths-related CPD, the teaching of mathematical reasoning concepts and any other maths-related programmes or interventions used in the class. In addition, this survey will use Likert scales to collect data on self-reported outcomes for teachers at baseline and endpoint, including an understanding of quantitative and numerical reasoning and the perceived importance of teaching this from KS1.

This information will help us understand the extent to which the programme differs from standard school practice and the extent to which the schools asked *not* to deliver the programme can legitimately be understood to represent a 'control' in terms of how similar their usual practice is to what the MR programme involves. This will inform our interpretation of the findings from the impact analysis.

Teacher & TA survey

All teachers and TAs in programme schools will be asked to complete an online survey at the end of the programme period. The survey will be routed to ensure that each individual only responds to questions relevant to their role. The survey will look to provide a broader perspective on the questions being addressed as part of the case studies, including staff completion and experience of the training (including whether the teacher and TA completed it together), support and delivery of the programme, key challenges and facilitators for

implementation, and perceived outcomes for the participating pupil and staff, as well as scope for spreading lessons learned to the broader school community. The survey will also cover the same questions on teacher self-reported outcomes as the BAU surveys to allow for pre-post comparisons between intervention and control.

All surveys will be hosted on Questback, and each participant will receive a unique link. Survey routing will deliver only relevant questions to minimise the completion burden for practitioners.

4. Monitoring information

Training course completion data

The Oxford University team will record completion of each of the training modules[23] and attendance at each of the webinars at school level and share this data with the NFER team in pseudonymised form. We will use these to assess fidelity in relation to training course completion.

Webinar attendance data

TLs will keep a record of attendance at each of the three webinars they deliver, and the Oxford University team will share this with NFER. This data will inform compliance and fidelity analysis.

Session delivery and attendance data

Each programme class will be asked to complete a session delivery log over the course of the delivery period where they recorded when each session was delivered, which programme units it covered, whether a TA was present and pupil attendance. This will inform our compliance and dosage analysis (see above), as well as our understanding of the extent to which the programme units were delivered in order and whether the programme delivery in practice aligned with the intended delivery schedule. Pupil attendance data will also be used to understand the programme reach and whether there were any trends in particular pupil groups missing programme sessions.

Computer games data

Back-end data from the computer games website will be extracted by the developer and shared with NFER in pseudonymised form. This will inform the impact and dosage analysis (see above), as well as our understanding of the extent of variation in the number of games played. We will also use the time stamps associated with when each game was played to descriptively assess the extent to which games were played outside of school time. We will also use statistical testing to determine whether the chance of a pupil playing games outside school time may be related to FSM status. Further exploratory analysis will be carried out as part of the impact section (see Footnote 9).

[23] A module is considered complete if the participant has submitted a response to at least one interactive activity.

Analysis

Qualitative data – observations, interviews & reviews

Interview notes will be written up as intelligent verbatim transcripts[24] and uploaded to the qualitative data analysis software MAXQDA. The data will then be analysed using qualitative content analysis, which looks to find and examine patterns of 'sense-making' through the content and underlying themes and meaning that emerge in a text (Biggs *et al.*, 2021).High-level deductive coding (approaching data with a pre-established framework for interpreting it) will be used to sort the data into relevant themes. Detailed inductive coding (identifying patterns of meaning present in the data) will then allow us to draw out the key findings under each of these themes.

Observation data will be treated qualitatively to provide a clear narrative of what the programme looks like in practice and the key variables that influence its effectiveness. Session observation notes will be typed up and uploaded to MAXQDA. We will interrogate the nature and prevalence of themes emerging in the data across the different case studies to draw out the key findings under each area of the observation tool across the sample as a whole. A separate matrix of the key findings from each data source for each case study will also facilitate within-case analysis. This means we will be able to triangulate the data sources and better understand the context within which particular issues or perspectives emerge.

All deductive coding will be based on a coding frame that will be developed in advance based on the programme ToC and IPE research questions to ensure relevance and minimise bias. To ensure our qualitative analysis is robust and consistent between different coders, a common approach will be discussed and agreed upon by the coding team in advance, and an initial subset of the data will be double-coded and cross-checked by the Project Manager. This will be facilitated by the use of MAXQDA qualitative analysis software, which creates a database of the codes assigned to each text. Further cross-checking will occur if and as required throughout the coding process. Any questions or uncertainties that emerge will also be addressed by the coding team as a whole.

Quantitative data – surveys and administrative data

As outlined above, survey response data will be exported from Questback and quality assured prior to its analysis, with each data source stored in a separate file.

The analysis of the surveys and MI will be designed and conducted by a Statistician in consultation with the Project Leader and Project Director. All quantitative analysis will be conducted using R. All codes, as well as the outputs, will be reviewed and checked by another experienced member of our Centre for Statistics. The required analyses will include both the descriptive statistics and inferential statistics required to answer the IPE research questions, including a small number of cross-tabulations for key potential moderators.

[24] Intelligent verbatim transcription excludes fillers and redundancies that do not add meaning to the content to make the text more 'readable' (McCullin, 2023).

Exploratory analysis of self-reported practitioner outcomes will statistically test for differences between a small number of single items at baseline and endpoint (selected based on their relevance to the ToC), comparing intervention and control groups using Mann-Whitney tests (treating the Likert-type scales as ordinal data). This will allow us to determine whether the change from baseline to endpoint for teachers in programme schools differs significantly from the same change for practitioners in control schools.

Triangulation of qualitative and quantitative data

The design of the data collection tools for each of the qualitative and quantitative components of the IPE will mutually inform each other (sequence permitting) to ensure consistency and create opportunities for complementary analysis. For example, similar questions will be asked in the case study interviews and Teacher and TA Survey, which will enable us to explore the details of a particular issue (in the interviews) as well as how perspectives on this may vary more broadly across the programme settings (in the survey). Emerging findings across the data collection activities will be logged by the researcher in a central log in real-time to ensure any relevant lessons feed into the design and delivery of future data collection activities.

We will also develop an integrated analysis framework that will map how each of the data sources will feed into our analysis and reporting for each of the IPE research questions. Findings from each of these data sources will subsequently be examined in tandem to ensure their integration when responding to each research question in the final report. We will collate and triangulate all data sources through an analysis workshop to ensure we provide a comprehensive assessment of the implementation effectiveness and perceived outcomes of the MR programme and inform our interpretation of findings from the IPE. A subsequent workshop with the broader evaluation team would allow for the IPE and IE findings to be brought together and explored within the context of the other by all members of the research team to understand, in particular, what further insights the IPE may lend to why or why not impacts may have been observed.

Reporting

The same researcher will carry out both the analysis and reporting for their allocated sections of the report. The analysis framework will ensure that each researcher has access to all the data relevant to their area(s) of focus. Once the data from each source has been transformed using qualitative codes or descriptive statistics, the researcher will draw out themes across the range of data types and use the unique insights each offers to provide a more integrated picture. Survey findings will, for example, provide an overview of the prevalence of particular experiences, while interview data will allow for more detailed illustrations of what this may look like in practice. Quantitative and qualitative data will be interleaved throughout the IPE section of the report, which will be structured according to the IPE dimensions and associated themes. Tables and figures will be used where they provide added value and clarity to the findings communicated through the text.

Cost evaluation design

A cost evaluation was carried out as part of the first effectiveness trial of MR (Stokes *et al.*, 2018). However, given the change in the training model, which constitutes a large part of the programme cost (in terms of school staff), it is important that a further cost evaluation is carried out to understand what this shift means for the relative affordability of the programme for schools. We will collect information on the pre-requisite, set-up and ongoing costs to schools of implementing the MR programme. Data collection for the cost evaluation will be embedded within the IPE activities to minimise burden on schools while enabling triangulation from various sources.

All case study schools will be asked to complete a detailed pro forma outlining their expenditures on training and programme delivery, including staff time. While the pro forma will be completed online during the session delivery period, teachers will be provided with a paper copy during the case study visits to facilitate preparation for completing the form.

In addition, the Teacher & TA Survey will include low-burden cost questions that are likely to have high variability across schools, including the extent to which pre-requisites such as screens and devices were usually already available for use and the range of staff time spent on various programme elements, such as the online training course, session preparation, webinar attendance and engagement with the community of practice. We will also look to understand the extent to which this time replaces activities in which they would have otherwise engaged (such as preparation for the lesson they would otherwise have delivered, or engagement with other maths CPD), or whether it is on top of their existing workload. We will likewise explore whether teachers and TA were given designated time by their school for programme activities and, if so, the extent to which this covered the time they ultimately spent on the programme. Time spent by teachers and TAs will be recorded separately to capture the different costs associated with different levels of seniority.

As the programme is intended to be delivered in place of usual teaching time, teacher time to deliver will not be included in the cost evaluation. However, TA time will be considered as part of the sensitivity analysis, using data collected from schools on the extent to which the TA would or would not have still been in the classroom for that period. We will also consider in descriptive terms any wider staffing implications this may have had for the school.

In order to establish relative cost compared to BAU, we will include relevant questions in the BAU surveys at baseline and endpoint. These questions will cover availability of the technological pre-requisites of the programme, time spent on preparation for maths lessons and maths-related CPD, and allocation practices around TA time.

As the programme is not currently commercially available, we will work with the Oxford University team to estimate the costs of implementing the programme, in accordance with EEF guidance. As the focus is on cost to schools, we will not consider the cost of TL training separately but ensure that it is accounted for in the amount that would be charged to schools to access the programme should it become commercially available with the current training model.

We will conduct the cost evaluation analysis in line with the EEF's latest cost evaluation guidance. Each cost to a school (e.g. photocopying materials) will be estimated per year, over a projected three-year period. Ongoing costs in years 2 and 3 will either be reduced to zero (for fixed costs) or we will make an informed decision about whether they are likely to change over time. Time and cost estimates will be reported in terms of means and ranges. Having established a cost per-school-per year, this figure will be divided by the total number of intervention pupils to estimate the cost per-pupil-per-year.

By default, we will assume that schools have the required technology in place to participate in MR, in particular that there are sufficient laptops/tablets and internet access for pupils to participate in the computer games. We will, however, relax this assumption in a sensitivity analysis, estimating costs for scenarios where schools need to purchase more laptops/tablets. If further costs emerge in the results that are both highly variable and represent a large proportion of total costs (on average), these will be included in further sensitivity scenarios.

We will also produce a table detailing the additional time commitments entailed by the programme for teachers and TAs (separately) on top of their regular workload. These time requirements will be broken down into training, preparation and delivery sections.

Ethics and registration

This evaluation will be conducted in accordance with NFER's Code of Practice. All of NFER's projects abide by its Code of Practice, which is in line with the Codes of Practice from BERA (the British Educational Research Association), MRA (the Market Research Association) and SRA (the Social Research Association), among others. NFER is committed to the highest ethical standards in all of its activities and ethical considerations are embedded in its detailed quality assurance processes. Every project is assessed against the NFER Code of Practice at proposal stage, with ethical approval a requirement of proceeding with the bid. Any significant updates to the project after this point are submitted to the Committee for further approval as needed.

Each participating school's headteacher will provide their agreement on behalf of the school to participate in the trial by signing the Memorandum of Understanding (MoU), which outlines the responsibilities of all parties involved in the trial.

NFER will share a parent letter and withdrawal form with schools to be sent to parents/carers of all pupils in participating classes. Through the withdrawal form, parents/carers will have the opportunity to withdraw their child from the evaluation and associated data processing at any stage of the trial.

A separate opt-in consent process will be used for the pupil focus groups and will only apply to those selected to participate. Given that pupils participating in this study are only 6 to 7-years-old, we cannot assume that all pupils will have the capacity to provide fully informed consent to participate. We will therefore provide parents/carers with a written information sheet about the focus groups which will contain full details about the focus group and what their child will be asked to do. Parents/carer will then be asked to provide written opt-in consent of their willingness for their child to be invited to participate in the focus group, by returning a consent form to the school, who will then pass this information on to the research team.

Pupil participation in the focus groups is voluntary, therefore even if a parent/carer has given consent for their child to participant, their child can still choose not to take part. Age-appropriate information about the focus groups will be provided to pupils at the same time as parents/carers receive information about the focus groups to allow them to discuss participation together. The researchers will also read this information to pupils at the beginning of the focus group to ensure pupils understand it and have the chance to ask any questions. If, at this point, a pupil decides that they would prefer not to participate, then they will be able to return to their class. Prior to beginning the focus group, the researchers will agree some ground rules for the group with the pupils and have a discussion with them about the types of scenarios in which we would need to break confidentiality, to ensure they fully understand what this means.

Interviewees (e.g. school staff) will be provided with information about the research and how we use their data before our visit and informed consent will be obtained from interviewees at the start of the interview If an individual staff member within a case study school does not wish to participate in the data collection they can choose to decline.

The trial will be designed, conducted and reported to CONSORT standards. It will be registered in the ISRCTN (International Standard Randomised Controlled Trial Number) Registry once the protocol has been finalised. The ISRCTN number will be added to this document as soon as it becomes available. The registry will be updated will the trial outcome upon its completion.

Data protection

All data gathered during the trial will be held in accordance with the data protection framework created by the Data Protection Act 2018 and the General Data Protection Regulation 2016/679 and will be treated in the strictest confidence by the NFER, the Oxford University team and the EEF. No individual or school will be identified in any report.

NFER is the data controller for evaluation and will make decisions about how and what personal data is used in accordance with the objectives of the study set by EEF. The University of Oxford is the data processor for the evaluation and data controller for the programme delivery.

The legal basis for processing personal data is covered by GDPR Article 6 (1) (f):

Legitimate interests: The processing is necessary for your (or a third party's) legitimate interests unless there is a good reason to protect the individual's personal data, which overrides those legitimate interests.

A legitimate interest assessment has been undertaken. The evaluation fulfils one of NFER's core business purposes (undertaking research, evaluation, and information activities). It has broader societal benefits and will contribute to improving the lives of learners by providing evidence about the impact of teaching techniques used in the classroom – in this case, the teaching of mathematical reasoning in the primary school classroom, and the relative effectiveness of the MR programme specifically.

The legal basis for processing pupils' special personal data (SEND status) is covered by GDPR Article 9 (2) (j) which states that *'processing is necessary for archiving purposes in the*

public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) (as supplemented by section 19 of the 2018 Act) based on domestic law which shall be proportionate to the aim pursued, respect the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject'.

We do not believe this processing will cause damage or distress to the pupils. The outcomes of the evaluation will not result in the creation of measures or decisions being made about individual pupils.

NFER and the Oxford University team will sign a Data Sharing Agreement (DSA) to govern the collection and sharing of personal data during this trial. This agreement includes a description of the nature of the data being collected and how it will be shared, stored, protected and reported by each party. In addition, the Oxford University team will provide a memorandum of understanding to schools, explaining the nature of the data being requested of schools, teachers and pupils, how it will be collected, and how it will be passed to and shared with NFER. Two separate Privacy Notices are available: one for schools and another one for parents.[25] All personal data will be shared via secure, password-protected data sharing portals.

The full name and contact details of the Headteacher will be collected when a school signs up to participate in the project. As part of the MoU, the full name, contact details and job role will be collected for the key contact on the project, who will facilitate communications between the school and the Oxford University team, and the school and NFER. The full name, role and contact details for the class teacher and TA in each participating class, will be collected for the evaluation to facilitate contact with participants. The Oxford University team will share pseudonymised Teacher and TA responses to questions asked as part of the online training course with NFER so that NFER can analyse rates of training completion. The surveys and interviews will ask about staff experiences of delivering the programme, challenges, and facilitators relevant to their school context and self-reported knowledge about and attitudes towards mathematical reasoning.

NFER will also collect pupil data from schools including names, date of birth, Unique Pupil Number (UPN), FSM eligibility status and PTM7 scores for all pupils in the participating classes. NFER will share pupil personal data with GL Assessment to enable the assessment marking process. For these pupils, background data including gender, FSM eligibility, EAL status and Special Education Needs & Disability (SEND) status will be collected from the National Pupil Database (NPD). To obtain the information from the NPD, NFER will securely provide the Data Sharing Team at the DfE with the names of the pupils, their dates of birth and UPNs, allowing a match to NPD.

[25] Both privacy notices are available here: https://www.nfer.ac.uk/for-schools/participate-in-research/participate-in-research-projects/effectiveness-trial-of-the-mathematical-reasoning-programme/

All NFER staff visiting schools will have up-to-date DBS checks. All data gathered during interviews will be stored securely. No names of individuals will be used in any report arising from this work.

Within three months of the end of project, NFER will send school and pupil data to EEF's data archive partner. At this point, EEF's data archive partner will keep a copy of the data and EEF will become the Data Controller. NFER will retain personal data for one year after report publication in case there are any queries about the report. One year after the report publication, all personal data will be securely deleted.

After the evaluation report has been published, NFER will share anonymised pupil data with the Oxford University team, who will then match this data with pseudonymised teacher data and school IDs that they will collect and hold.

NFER will not store or transfer any data outside of the UK. When we use Questback to administer online surveys, data is stored in the EU. GL Assessment may transfer personal data outside of the UK. However, this is safeguarded by the appropriate contractual safeguard.

Personnel

Table 4: Project team

Name	Organisation	Role and Responsibilities
Evaluation team		
Helen Poet	NFER	Project Director – responsible for overall delivery of the evaluation to agreed specifications and overseeing the integration of the impact and IPE. She will also be responsible for strategic oversight and the quality of the outputs.
Lillian Flemons	NFER	Trial Manager & IPE Lead – day-to-day management of the trial, and design and delivery of the IPE.
Andrew Smith	NFER	Impact Evaluation Design Lead – responsible for the design of the trial and analytical considerations, responsible for the protocol, study plan and quality assurance of the impact analysis.
Eleanor Bradley	NFER	IPE Researcher – IPE data collection, analysis and reporting
Gustavo Lopes	NFER	IPE Researcher – IPE data collection
Chris Morton	NFER	Statistician – contributing to the analytical design and running the quantitative analyses for the impact and IPE strands.
Kathryn Hurd	NFER	Research Operations Lead – responsible for leadership and strategy around data collection and school communications
Katharine Stoodley	NFER	Operations Manager – day-to-day operations, including preparation of recruitment documents, coordinating data

		collection and point of contact for schools participating in the trial
Delivery team		
Professor Gabriel Stylianides	University of Oxford	Principal Investigator and co-designer of the online professional development training for teachers – responsible for strategic oversight of, and contributor to, all aspects of delivery and related outputs
Professor Terezinha Nunes	University of Oxford	Programme designer, co-designer of the online professional development training for teachers and PI – contributes to all aspects of delivery, including academic, technical and administrative, and to writing and reviewing documents for programme implementation and evaluation.
Louise Matthews	University of Oxford	Research Project Manager responsible for the day-to-day management of the delivery, including the professional development of the Teacher Leaders.

Risks

Table 5: Project risks

No.	Risk	Risk Assessment		Mitigation/Counter Measures/Contingencies
		Likelihood	Impact	
1	Monitoring information (MI) data requires additional cleaning	Likely	Low	Close collaboration between NFER and the Oxford University team to agree a specification for MI data in advance. Update of the MI analysis plan once it is clear what data is available. If necessary, low quality/low completeness of data flagged to the EEF at the earliest opportunity.
2	Schools do not complete session delivery log	May Happen	Significant	During set-up we will establish the minimum data that needs to be collected in the log to minimise burden. NFER will work with the Oxford University team to design a manageable instrument and completion process. Our operations team will support schools with completion and be on hand to answer queries.
3	Insufficient number of schools recruited to the trial	May Happen	Significant	NFER to input into recruitment material and work closely with the Oxford University team throughout the recruitment process. If required, our experienced operations team can assist with recruitment through a separate grant agreement.

				Decide and monitor pre-agreed recruitment targets to identify any unfavourable trends early on to act quickly. Efficient and flexible approach to school and pupil data collection to allow for the recruitment window to be open for as long as possible.
4	School and pupil attrition from trial and primary analysis	Unlikely	Significant	Schools sign up for the trial via a Memorandum of Understanding with a clear identification of requirements. Clear initial and ongoing communication via one key contact per school explaining principles and expectations. We will keep them informed of upcoming activities, timelines and next steps and provide support on all activities to ensure that activities are completed. Support webinar offered during baseline data collection to allow practitioners to ask any questions. NFER Test Administrators to administer endpoint tests, at a convenient time for the school.
5	Difficulty in securing target response rates for IPE	May happen	Moderate	Communication with schools explaining research benefits. Ongoing reminders. Flexibility in timings of school visits and interviews. Close liaison with the delivery team to support IPE engagement. Online data collection (including remote interviews) where possible to minimise burden. 'Thank you' payments of £100 for all case study schools.
6	Impact estimates are biased as schools allocated to control arm adapt their teaching practices to produce similar outcomes	May happen	Moderate	Survey of schools in both trial arms to understand usual practice. Guidance for schools allocated to the control arm (i.e. to continue with their usual teaching practice).
7	Changes to the project	Likely	Low	NFER has a large research department with numerous experienced researchers and research who could be redeployed.

	team due to sickness, absence or staff turnover			Clear and accurate project documentation would support continuity in the event of any team changes.

Timeline

Table 6: Project timeline

Dates	Activity	Staff responsible/ leading
Dec 2023-Jan 2024	IDEA workshop and set-up meetings Development of recruitment documents	NFER & Oxford University
Feb-Jun 2024	School recruitment Teacher Leader Training	Oxford University
Jul 2024	Publication of protocol, trial registration and NPD application Pupil data collection from schools	NFER
Sept 2024	Baseline assessment Baseline BAU survey	NFER
Oct 2024	Randomisation Share pre-populated session delivery logs with schools	NFER
Nov 2024	Launch event, training and first webinars Training data collection	Oxford University NFER
Dec-Mar 2025	Programme delivery IPE case study data collection (*t1*), Teacher Leader interviews and webinars observations	Oxford NFER
Apr 2025	Publication of the Statistical Analysis Plan (SAP)	NFER
Jun 2025	Endpoint assessment Endpoint surveys IPE case study data collection (*t2)*	NFER
Jul 2025	IPE incentive payments Control school payments	NFER Oxford University
Sept 2025	Endpoint assessment data shared with schools	NFER
Dec 2025	Submission of draft report	NFER
Mar 2026	Publication of final report	NFER

References

Bishenden, O. (2023) Email to Andrew Smith, 17 November.

Biggs, R., De Vos, A., Preiser, R., Clements, H., Maciejewski, K. and Schlüter, M. (2021) *The routledge handbook of research methods for social-ecological systems*. London: Routledge. Available at: https://www.routledge.com/The-Routledge-Handbook-of-Research-Methods-for-Social-Ecological-Systems/Biggs-deVos-Preiser-Clements-Maciejewski-Schluter/p/book/9781032020761 (Accessed: 24 April 2023).

Ching, B.H.-H., Kong, K.H.C., Wu, H.X. and Chen, T.T. (2020) 'Examining the reciprocal relations of mathematics anxiety to quantitative reasoning and number knowledge in Chinese children', *Contemporary Educational Psychology*, 63, p. 101919. Available at: https://doi.org/10.1016/j.cedpsych.2020.101919.

Ching, B.H.-H. and Nunes, T. (2017) 'The importance of additive reasoning in children's mathematical achievement: a longitudinal study.', *Journal of Educational Psychology*, 109(4), pp. 477–508. Available at: https://doi.org/10.1037/edu0000154.

Çiftçi, S.K. and Yildiz, P. (2019) 'The effect of self-confidence on mathematics achievement: the metaanalysis of Trends in International Mathematics and Science Study (TIMSS)', *International Journal of Instruction*, 12(2), pp. 683–694. Available at: https://doi.org/10.29333/iji.2019.12243a.

Henry, L. (2015) 'The effects of ability grouping on the learning of children from low income homes: a systematic review', *The STeP Journal*, 2(3), pp. 79–87.

Hewitt, C.E. and Torgerson, D.J. (2006) 'Is restricted randomisation necessary?', *BMJ*, 332(7556), pp. 1506–1508. Available at: https://doi.org/10.1136/bmj.332.7556.1506.

Johnston, O. and Wildy, H. (2016) 'The effects of streaming in the secondary school on learning outcomes for Australian students. A review of the international literature', *Australian Journal of Education*, 60(1), pp. 42–59. Available at: https://doi.org/10.1177/0004944115626522.

Jordan, N.C., Devlin, B.L. and Botello, M. (2022) 'Core foundations of early mathematics: refining the number sense framework', *Current Opinion in Behavioral Sciences*, 46, p. 101181. Available at: https://doi.org/10.1016/j.cobeha.2022.101181.

Nunes, T., Bryant, P., Barros, R. and Sylva, K. (2012) 'The relative importance of two different mathematical abilities to mathematical achievement', *The British Journal of Educational Psychology*, 82(1), pp. 136–56. Available at: https://doi.org/10.1111/j.2044-8279.2011.02033.x.

Nunes, T., Bryant, P., Evans, D., Bell, D., Gardner, S., Gardner, A. and Carraher, J. (2007) 'The contribution of logical reasoning to the learning of mathematics in primary school', *British Journal of Developmental Psychology*, 25(1), pp. 147–166. Available at: https://doi.org/10.1348/026151006X153127.

Parsons, S. and Hallam, S. (2014) 'The impact of streaming on attainment at age seven: evidence from the Millennium Cohort Study', *Oxford Review of Education*, 40(5), pp. 567–589. Available at: https://doi.org/10.1080/03054985.2014.959911.

Putwain, D.W., Becker, S., Symes, W. and Pekrun, R. (2018) 'Reciprocal relations between students' academic enjoyment, boredom, and achievement over time', *Learning and Instruction*, 54, pp. 73–81. Available at: https://doi.org/10.1016/j.learninstruc.2017.08.004.

Singh, A., Uwimpuhwe, G., Vallis, D., Akhter, N., Coolen-Matur, T., Higgins, S., Einbeck, J., Culliney, M. and Demack, S. (2023) *Improving power calculations in educational trials*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/methodological-research-and-innovations/Work_Package_2023-WP6_18_09_2023_FINAL.pdf?v=1696410358 (Accessed: 20 March 2024).

Stokes, L., Hudson-Sharp, N., Dorsett, R., Rolfe, H., Anders, J., George, A., Buzzeo, J. and Munro-Lott, N. (2018) *Mathematical reasoning: evaluation report and executive summary*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Mathematical_Reasoning.pdf?v=1696414461 (Accessed: 20 March 2024).

Worth, J., Sizmur, J., Ager, R. and Styles, B. (2015) *Improving numeracy and literacy: evaluation report and executive summary*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Oxford_Numeracy_and_Literacy.pdf?v=1696414466 (Accessed: 20 March 2024).

Appendix A: Changes since the previous EEF evaluation

Feature		Efficacy to effectiveness stage	First effectiveness trial to second effectiveness trial
Programme	Programme content	No change.	No change.
	Delivery model	One day of in-person training and coaching delivered by Oxford University developers changed to a train-the-trainer model. Oxford University trained Work Group Leaders on two days of in-person training, who then provided one in-person training day and one coaching day to teachers.	Change from an in-person train-the-trainer model to one and a half days of training made up of an online course offered by the team of developers, three webinars, an online forum and support via email, if required, from Teacher Leaders. Teacher Leaders receive three days of training from the Oxford University team, in a blended format (some online elements and some in-person training).
	Programme duration	Schools were advised to deliver the programme's 10 units over 12-15 lessons.	No change.
Evaluation	Eligibility criteria	No change.	Change from only schools within the eight Maths Hubs to all English state primary schools.
	Level of randomisation	Randomisation within blocks was defined based on hubs, the proportion of children eligible for FSM, and prior attainment at KS1.	Simple randomisation without stratification.
	Outcomes and baseline	The baseline and endpoint have been changed to an updated version of Progress in Maths (PiM) 7—Progress Test in Maths (PTM) 7.	No change.
	Control condition	No change.	No change.

Appendix B: GL Assessment Mathematics process categories

FF: Fluency in facts and procedures

Pupils can, for example:

- recall mathematical facts, terminology and definitions (such as the properties of shapes);
- recall number bonds and multiplication tables;
- perform straightforward calculations.

FC: Fluency in conceptual understanding

Pupils can, for example:

- demonstrate understanding of a mathematical concept in the context of a routine problem (for example, calculate with or compare decimal numbers, identify odd numbers, prime numbers and multiples);
- extract information from common representations, such as charts, graphs, tables and diagrams;
- identify and apply the appropriate mathematical procedure or operation in a straightforward word problem (for example, knowing when to add, multiply or divide).

MR: Mathematical reasoning

Pupils can, for example:

- make deductions, inferences and draw conclusions from mathematical information;
- construct chains of reasoning to achieve a given result;
- interpret and communicate information accurately.

PS: Problem-solving

Pupils can, for example:

- translate problems in mathematical or non-mathematical contexts into a process or a series of mathematical processes;
- make and use connections between different parts of mathematics;
- interpret results in the context of the given problem;
- evaluate methods used and results obtained;
- evaluate solutions to identify how they may have been affected by assumptions made.

Appendix C: MR Programme 'keys to success' for teachers

The list of Keys to success below includes examples of what teachers do as they aim to implement the keys to success. Some examples fit in with more than one item in the list.

1. The children should always be actively solving problems. Each child should produce an answer for every problem. They should only discuss the answer after everyone has answered the question.

   Teachers use prompts and questions to keep the children thinking, explaining and discussing their answers; they allow time for individual thinking; they guide the children to use the extension activities if the children finished earlier and use paired discussion.

2. They should have manipulatives available to help them all the time. They can use these when they need them even when in our booklet we do not specifically state that they should use these materials.

   Teachers demonstrate relations between quantities and between numbers using manipulatives; they encourage the children to use manipulatives to explain their thinking and to demonstrate relations between quantities and numbers; they make sure that the children have manipulatives available all the time.

3. Discussing their answers is an important part of learning. Both children who have right and wrong answers need an opportunity to demonstrate their thinking. Acting out word problems with materials helps them to connect the word problems with real life and helps them to show their reasoning.

   Teachers consistently ask at least two children to explain their reasoning; they value verbal explanations as well as practical demonstrations using manipulatives.

4. Even when all the children have made mistakes, we don't need to tell them the answer. The teacher can scaffold the solution: i.e. create a support for the children to think again.

   Teachers start a solution with the materials and then see whether the children can continue the process; they support the children's reasoning with further questions. For example, in a start-unknow problem they ask *'How many sweets did the girl have to begin with?'*. If the children are not able to make a start, they might say '*Show me the sweets the girl has now; show me how many of these her granny gave her; what about these other sweets, where did they come from*?'

5. We try to create opportunities for children to learn to be flexible in their use of language. For example, the word "more" is not always connected to addition: *Anna has 8 sweets and Sharon has 5. How many more sweets does Anna have than Sharon?* The children need to learn to use language flexibly because this is going to be part of their future learning too.

   Teachers use the language in the PowerPoint and sometimes rephrase it too, after the children give their answer; they verbalise the solutions that the children produce in action with manipulatives: for example, "*I see, you place 3 sweets in front of each plate*, *so you could find out how many sweets altogether*".

6. The aim of our activities is to encourage children to reason about relations between quantities and between operations. In the activities about the inverse relation between addition and subtraction, they should be able to reason that if you add and take away the same number of bricks to a row, the number in the row does not change. This is why we let them count the bricks in the initial row and then cover it.

   Teachers encourage the children to think about what happens when you undo an action: for example, when you add some blocks and then take the same blocks away; when you add some blocks and take an extra one away; teachers ask the children to think using further questions: for example, if you know how many sweets the girl had after her granny gave her five, how can you get back to the number she had before her granny gave her these five?

7. In the word problems, they will be learning how to use counting in different ways to solve different types of problems. Here the logic that they see in the situation should guide the way that they count. It is only later that they will connect these different ways of counting with calculating, but if they don't understand the logic of these situations, they will not understand the arithmetic operations later on.

   Teachers emphasise the logic of part-whole in additive reasoning by asking the children to think about the parts and the whole as they count; for example, they ask the child show me the marbles the boy the boy still had when he got home; what about the others, what happened to these marbles? Teachers emphasise the one-to-many correspondence in multiplicative reasoning: the crucial wording relates to a fixed ratio, i.e. how many *x* for each *y*; this is the same for multiplication and division.

Appendix D: Training and preparation for Teacher Leaders

The training of TLs blends in-person and online activities. All TLs receive three days of in-person training with the Oxford University team. The first in-person training session provides an introduction to the programme and the online training course and encourages reflection and discussion around their own delivery of the programme sessions for practice (see below). It also includes discussion of themes that TLs will need to address when running webinars, such as why teach mathematical reasoning, and watching videos in order to identify observed "keys to success". These videos are part of the CPD for teachers in Module 5. The second training session prepares TLs for delivering a launch event and the three webinars, as well as how to support teachers and TAs with the training process. The third session focuses on feedback and reflection. New TLs also attend webinars delivered by the Oxford University team. These are different from the webinars the TLs themselves deliver to the teachers and TAs as part of the programme and are designed to support the Teacher Leaders with their own teaching of the programme.

All TLs are required to complete an online course tailored to their role. The course introduces TLs to the concept of mathematical reasoning, the expectations of the TL role, how to support teachers during their training (including through promoting fidelity) and how to deliver each of the three webinars. Participants are asked to respond to regular interactive activities to encourage them to reflect on what they have learnt and how they will apply this in practice. The course comprises training videos, research briefs, an online forum, interactive elements, and drafts of emails and other documents that TLs will be required to send to schools. TLs receive additional training and support material and templates for the programme as part of the in-person training session, as well as a slide deck for each of the webinars they deliver to the teachers and TAs as part of the programme. They are also provided with responses to FAQs from schools.

All TLs are expected to complete the online training course for schools and practise delivering eight sessions of the programme in non-trial schools. They will also shadow delivery of the school support webinars (see below) by an experienced TL before delivering their own.[26]

All training and preparation activities except the third day of in-person training and the webinar shadowing must be completed prior to in-school implementation of the programme.

[26] In the context of this trial, three of the nine TLs have already had experience in the role through participating in the pilot study for the online training approach.